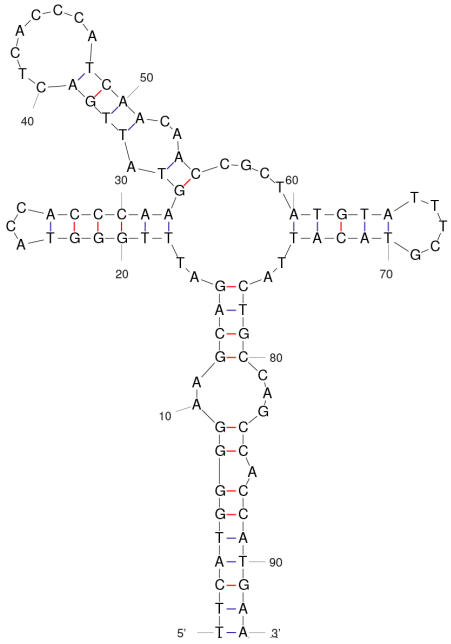


In vitro detection of non-canonical mitochondrial DNA conformations

João Pedro Silva Monteiro
Mestrado em Genética Forense
Departamento de Biologia
2014/2015

Orientador
Filipe Pereira, PhD, Investigador no Centro
Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR)

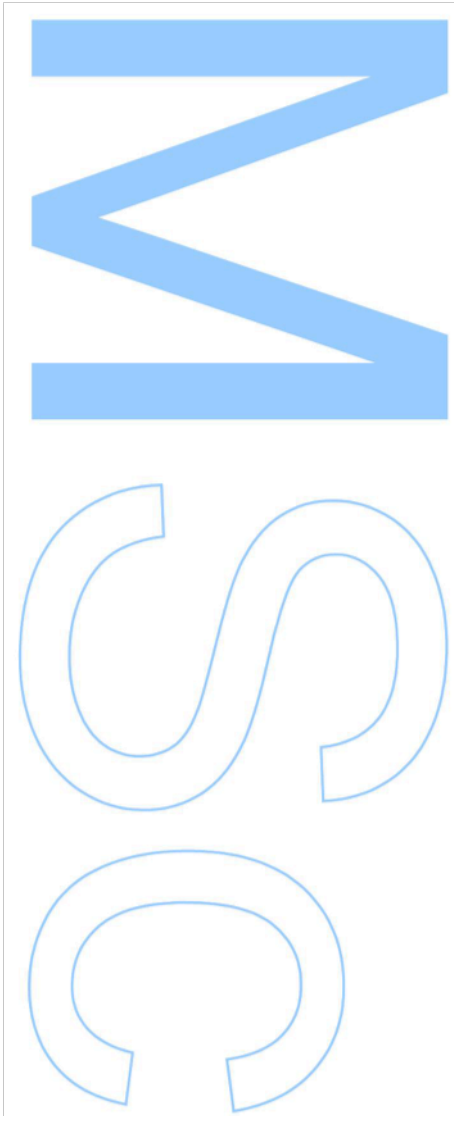
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



Dissertação de candidatura ao grau de Mestre em Genética Forense, submetida à Faculdade de Ciências da Universidade do Porto.

O presente trabalho foi desenvolvido no Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR), sob a orientação científica do Doutor Filipe Pereira.

Dissertation for applying for a Master's degree in Forensic Genetics, submitted to the Faculty of Sciences of the University of Porto.

The present work was developed at the Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), under the scientific supervision of Filipe Pereira, PhD.

“I can’t change the direction of the wind, but I can adjust my sails to always reach my destination.”

Jimmy Dean

Agradecimentos

Gostaria de agradecer ao Filipe pela oportunidade que me deu para desenvolver este projeto, pelo apoio que me deu, que nos permitiu ultrapassar os obstáculos que se iam colocando ao longo deste percurso e continuar com este trabalho em frente. Nunca esquecendo o serviço de táxis... Temos que ver como ficou a tarifa!

Um obrigado também a todos aqueles com quem partilhei o laboratório ao longo deste ano, os fisces. O vosso companheirismo permitiu manter a sanidade mental em momentos em que nem um simples gel queria polimerizar. E que venham mais convívios para festejar tudo e mais alguma coisa!

A vocês meus amigos, minha malta do costume, um enorme obrigado! O vosso apoio foi fundamental, tal como as conversas, as gargalhadas... todos aqueles momentos incríveis (com ou sem carta registada prévia) que partilhámos juntos! Nada disto tinha feito sentido sem vocês.

Agradeço também a todos aqueles que de uma forma ou de outra também me ajudaram e incentivaram ao longo deste ano. Em especial à Sofia, cujo o apoio e força foram essenciais para que, mesmo nos momentos mais difíceis, eu seguisse em frente. Obrigado por tudo!

Aos meus pais, obrigado por tudo. Obrigado por me terem dado todas as oportunidades, por me permitirem crescer fora de casa, por me deixarem seguir o caminho que escolhi para mim. Espero orgulhar-vos!

À minha irmã Joana, agradeço-te por seres um grande exemplo para mim, por me mostrares que com força de vontade e perseverança não há barreiras que nos impeçam de seguir aquilo que queremos! Paulo, obrigado também por todas as conversas que partilhámos, que me permitiam pensar e ver mais além. E a ti Vasco, quando um dia leres isto por engano, obrigado por teres ajudado imenso o tio, sempre a querer brincar com o meu portátil enquanto eu trabalhava!

Resumo

As cadeias complementares do DNA mitocondrial (mtDNA) humano estão organizadas quase sempre na sua forma canónica em dupla hélice. O nosso conhecimento em relação às conformações alternativas do mtDNA e as suas interações com o ambiente mitocondrial permanece bastante limitado. É plausível que muitas alterações à forma canónica do DNA ocorram no genoma mitocondrial e com consequências biológicas importantes, como já foi demonstrado inequivocamente noutros sistemas genéticos.

Recentemente, vários estudos demonstraram que estruturas secundárias de DNA (muitas vezes apelidadas de não canónicas, alternativas ou estruturas secundárias de DNA) ocorrem, pelo menos transientemente, no genoma mitocondrial. Neste estudo testámos *in vitro* a possível formação de conformações secundárias no mtDNA. Usámos uma estrutura com 93 nucleótidos prevista anteriormente (chamada de estrutura A) para desenhar oligonucleótidos que incluíam sequências capazes de formar *hairpins*. Também desenhámos oligonucleótidos com uma sequência que incluía a origem de replicação da cadeia leve (O_L) e uma sequência do plasmídeo ColE1, ambas bem conhecidas pelo potencial de adotar conformações não canónicas *in vivo*. Ainda foram usadas duas regiões do mtDNA com baixo potencial de formação de estruturas, ND2 e ND3, como controlo para as experiências, uma vez que estas sequências devem formar cadeias duplas perfeitas. Os *constructs* foram sujeitos a digestão com uma endonuclease (nuclease S1) que corta nucleótidos não emparelhados nos *hairpin loops*. Demonstrámos que, individualmente, dois dos três *hairpins* que constituem a estrutura A e o seu *stem* têm o potencial para adotar estruturas secundárias *in vitro*, apoiando a possibilidade de formação da estrutura A. Esta conclusão é suportada pelos resultados obtidos com o Construct_O_L, ND2 e ND3. Testamos também um procedimento para a observação dos *constructs* em cadeia dupla de DNA através de microscopia de força atómica (AFM). Embora não tenha sido possível observar as estruturas, os nossos resultados irão servir como guia para experiências futuras sobre o assunto.

Preparámos uma base de dados onde identificámos 579 posições heteroplásmicas no mtDNA. Os nosso resultados apontam para uma maior presença de posições heteroplásmicas na região controlo, comparando com o mtDNA

completo. Descobrimos que tanto no mtDNA completo ou apenas na região de controlo a localização de heteroplasmias é mais comum em estruturas secundárias. Para além disso, no mtDNA completo, as heteroplasmias são mais propícias nos *stems* do que em nucleótidos desemparelhados.

Em relação à estrutura A, anotámos um total de 12 posições heteroplásmicas, sem qualquer diferença estatisticamente significativa entre as suas localizações. Apenas duas não provocaram qualquer alteração estrutural comparando com a estrutura A prevista, 16043G e 16093C. O facto da posição 16093 ter sido repetidamente reportada com uma alta percentagem de heteroplasmia pode ser devido ao facto da estrutura A não ser modificada com a mutação que ocorre. Desta foram, qualquer papel regulador que esta estrutura possa desempenhar mantem-se intacto. As mutações que provocam alterações estruturais podem ser observadas em percentagens inferiores, possivelmente porque, atingindo um determinado limiar, poderão ter efeitos patológicos desconhecidos.

Palavras-chave: mtDNA, estruturas secundárias de DNA, estrutura A, detecção *in vitro*, oligonucleótidos, endonuclease, AFM, heteroplasmia, distribuição de posições heteroplásmicas, alterações estruturais, posição 16093.

Abstract

The complementary strands of the human mitochondrial DNA (mtDNA) are primarily organized into the canonical right-handed double-helical structure of B-form DNA. Our knowledge of the higher order topology of mtDNA and its interaction with the mitochondrial environment remains quite limited. It is plausible that many deformations to the canonical B-form of DNA occur in the mitochondrial genome and have important biological consequences, as has been unequivocally shown in other genetic systems.

In recent years, several studies have demonstrated that non-B DNA structures (often called non-canonical, unusual, alternative or secondary DNA structures) occur, at least transiently, in the mitochondrial genome. In this study we tested *in vitro* the putative formation of non-B DNA conformations in mtDNA. We used a predicted 93-nucleotide cloverleaf-like non-B DNA structure (named structure A) to design oligonucleotides able to form the hairpins predicted for this structure. We also designed oligonucleotides with a sequence that included the origin of replication of the L-strand (O_L) and a sequence from the plasmid ColE1, both known to adopt non-canonical conformations *in vivo*. We also used two regions of mtDNA with low folding potential, ND2 and ND3, as controls for the formation of non-B DNA, since these sequences should form perfect duplexes. The constructs with the hairpins were subjected to digestion with an endonuclease (S1 nuclease), which cleaves unpaired nucleotides at the hairpin loops. We demonstrated that, individually, two of the three hairpins that constitute the cloverleaf-like structure A and its stem have the potential to adopt secondary structures *in vitro*, supporting the possible formation of structure A. This finding was supported by the results obtained with the controls Construct_O_L, ND2 and ND3. Here we also describe a procedure for observation of duplex DNA constructs by atomic force microscopy (AFM). Although it was not possible to observe the structures, our findings will serve to guide future experiments on the subject.

We prepared a database where we identified 579 published mtDNA heteroplasmic positions. Our results demostrate a higher presence of heteroplasmic sites in the control region when compared to the entire mtDNA. We also found that heteroplasmic positions are more common within secondary structures. Moreover, the stems are more prone to heteroplasmy than unpaired nucleotides.

Regarding structure A, a total of 12 heteroplasmic positions were noted, without any statistical significant difference between their locations. Only two of them did not cause structural alterations compared to the predicted structure A, 16043G and 16093C. The fact that position 16093 has been repeatdly reported with a high percentage of heteroplasmy may be due to the fact that the structure A is not modified with the mutation that occurs. Therefore, any potential regulatory role that this structure may play remains unchanged. The mutations that cause structural alterations may be observed in lower percentages, because, possibly, if they reach a certain threshold, they could have any unknown pathological effect.

Keywords: mtDNA, non-B DNA, structure A, *in vitro* detection, oligonucleotides, endonuclease, AFM, heteroplasmy, heteroplasmic positions distribution, structural alteration, position 16093.

Table of contents

List of tables

List of figures

List of abbreviations

A	Adenine
ATP	Adenosine triphosphate
APS	Ammonium persulfate
AFM	Atomic force microscopy
bp	Base pair(s)
CRS	Cambridge Reference Sequence
° C	Celsius degree
CR	Coding region
CSBs	Conserved sequences blocks
C	Cytosine
DNA	Deoxyribonucleic acid
D loop	Displacement loop
dH_2O	Distillated water
dsDNA	Double stranded DNA
EDTA	Ethylenediaminetetetraacetic acid
e.g.	*exempli gratis*
ΔG	Free energy
G	Guanine
H-strand	Heavy strand
HPLC	High-performance liquid chromatography
HSP	H-strand promoter
HVI	Hypervariable region I
HVII	Hypervariable region II
HVIII	Hypervariable region III
i.e.	*id est*
indel	Insertion/deletion
L-strand	Light strand
LSP	L-strand promoter
$MgCl_2$	Magnesium chloride
mRNA	Messenger RNA
µg	Microgram
µL	Microliter
µM	Micromolar

mL	Milliliter
mm	Millimeter
mtDNA	Mitochondrial DNA
ng	Nanogram
NGS	Next-generation sequencing
nt	Nucleotide
O_H	Origin of replication of the H-strand
O_L	Origin of replication of the L-strand
OXPHOS	Oxidative phosphorylation
pmol	Picomol
PCR	Polymerase chain reaction
rCRS	Revised Cambridge Reference Sequence
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
ssDNA	Single stranded DNA
Na_2CO_3	Sodium carbonate
NaCl	Sodium chloride
TAS	Termination-associated sequence
TEMED	Tetramethylethylenediamine
kb	Thousands base pairs
T	Thymine
tRNA	Transfer RNA
tRNAPhe	Phenylalanine tRNA
tRNAPro	Proline tRNA
TEN	Tris-HCl, EDTA, NaCl
Tris-HCl	Trizma hydrochloride
V	Volts

1. Introduction

1.1. Mitochondria

Mitochondria are cytoplasmic organelles responsible for the production of adenosine triphosphate (ATP) through oxidative phosphorylation providing energy essential for life. It is now well established that mitochondria are also involved in essential processes for the cell, playing an important role in cell signaling pathways, regulating cell metabolism, development, division and death (McBride et al., 2006; Saccone, 2005). These semi-autonomous organelles possess their own genetic material, mitochondrial DNA (mtDNA), whose gene products are involved in mitochondrial oxidative phosphorylation, transcription and translation (Falkenberg et al., 2007). In recent years, a consensus has emerged from molecular phylogenetic investigations favoring the idea that all mitochondria had a common origin around two billion years ago. The mtDNA is a remnant of their origin in an α-protobacterium as explained by the widely accepted endosymbiotic theory, according to which the origin of the eukaryotic cell with mitochondria resulted from a symbiosis between an α-protobacterium and a "proto"-eukaryote (Bernt et al., 2013b; Thorsness and Hanekamp, 2001).

Although the genetic role of mtDNA seems to be conserved among different taxonomic groups, it presents noticeable differences regarding size, conformation and, also, in gene composition and arrangement. The importance of mtDNA is shown by its maintenance in the eukaryotic cell in very different phylogenetic lineages (Gray et al., 1999; Saccone, 2005).

1.2. Mitochondrial DNA

The human mtDNA is a double stranded circular DNA molecule of about 16.5 kilobases (kb) organized in nucleoprotein complexes (nucleoids) dispersed throughout the mitochondrial matrix, each one consisting of several mtDNA molecules (Chen and Butow, 2005; Legros et al., 2004). The two mtDNA strands are historically designated as heavy- (H) and light- (L) strands based on their different buoyant densities in a cesium chloride gradient resulting from a strand bias in base composition (the H-strand is guanine rich, whereas the L-strand is guanine poor).

The first complete human mtDNA sequence was reported in 1981 (Anderson et al., 1981), being known as Cambridge Reference Sequence (CRS). In 1999, the sequence was reanalyzed and corrected, being known as revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999). The rCRS is now used as reference for numbering positions of mtDNA. In addition to coding for the 13 essential membrane proteins of the oxidative phosphorylation (OXPHOS) machinery, the mtDNA also encodes two rRNA genes (12s and 16s rRNAs) and 22 tRNAs required for translation of mtDNA-encoded mRNAs in the mitochondrial matrix. The mitochondrial genome is characterized by a remarkable organization economy with 37 genes close to each other with no intronic segments interrupting the coding sequence, most of them encoded in the H-strand (**Figure 1**) (Taanman and Williams, 2005).

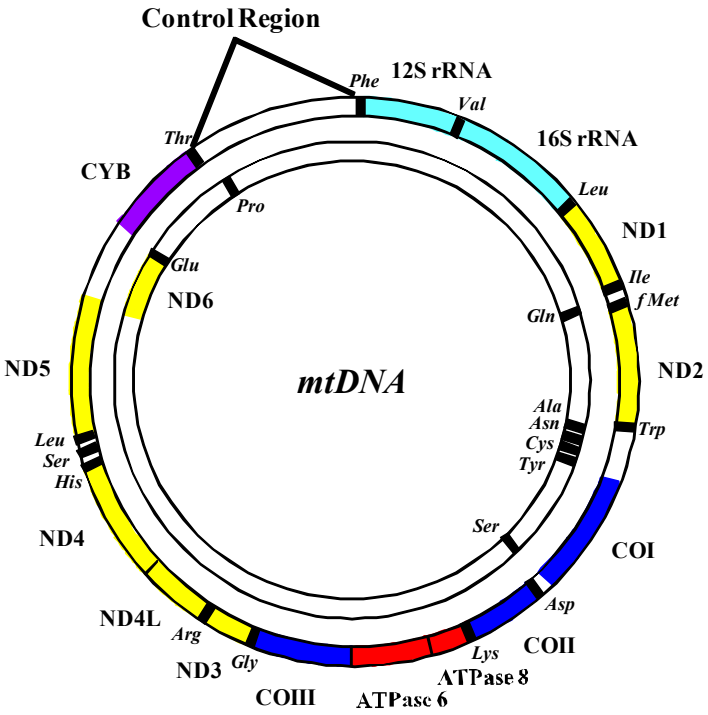


Figure 1 – Schematic representation of the human mtDNA molecule with the heavy- (H) and light- (L) strand on the outside and inside of the scheme, respectively. The mtDNA-encoded genes are colour coded as follows: rRNAs, light blue; complex I/NADH dehydrogenase, yellow; complex III/cytochrome b, purple; complex IV/cytochrome c oxidase, dark blue and complex V/ATP synthase, red. Black boxes labelled denote tRNAs for specific amino acids according to a three-letters nomenclature.

Intergenic domains are very rare and usually small, with the only exception being the control region (CR), also known as D-loop or AT-rich region (**Figure 2**). This noncoding segment is responsible for the precise binding of several nuclear-encoded proteins that regulate mtDNA replication and transcription. This region is

characterized by a peculiar genetic organization with highly conserved regions interspersed with some of the most quickly-evolving nucleotide positions identified so far. Although it only represents 7% of the mitogenome, the CR contains 25% of the variable sites (Bernt et al., 2013a; Krjutskov et al., 2014; Saccone, 2005). Indeed, certain nucleotide positions within the human mtDNA have been reported to mutate at rates that are significantly higher than average (mutational hotspots), a fact that has important implications for both evolutionary and phylogenetic studies (Brown et al., 1979; Krjutskov et al., 2014; Saccone, 2005). The mtDNA has a mutation rate higher that nuclear DNA due to fewer DNA repair mechanism, lack of histones (Krjutskov et al., 2014) and the exposure to reactive oxygen species (ROS), which cause oxidative damage (Bernt et al., 2013a; Butler, 2012).

The maintenance of conserved segments is believed to result from selective constraints acting on putative regulatory domains (Sbisa et al., 1997). Candidates for protein-binding regulatory signals are two classes of conserved segments identified by sequence comparisons: termination-associated sequences (TAS) located at the control region 5' end and conserved sequences blocks (CSBs) proposed to control the H-strand replication initiation on the control region 3' domain (Walberg and Clayton, 1981).

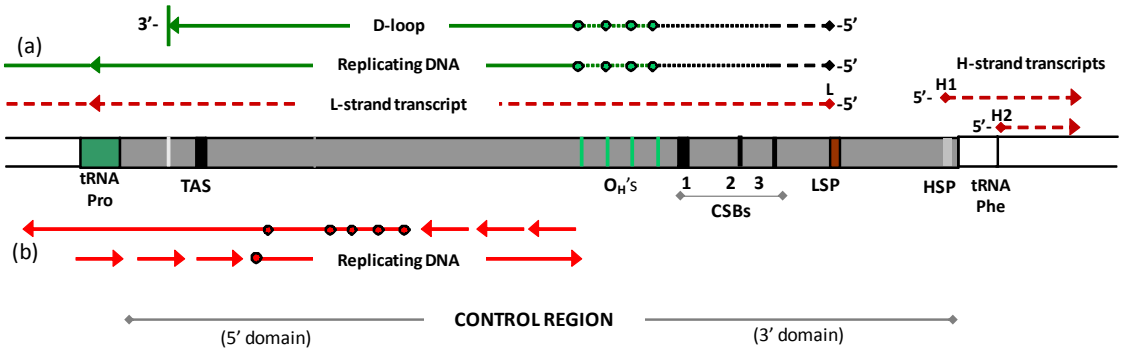


Figure 2 – Schematic representation of the mammalian mtDNA control region. Orientation and sites for replication initiation are indicated according to the strand-asynchronous (a) and the bidirectional strand-coupled (b) models. Localization of the phenylalanine transfer RNA (tRNAPhe), proline transfer RNA (tRNAPro), H-strand promoter (HSP), L-strand promoter (LSP), strand-asynchronous origin of H-strand replication sites (O_H), conserved sequence blocks (CSBs) and termination-associated sequences (TAS) are shown.

According to the asymmetric model of replication of mtDNA (**Figure 3**), the replication begins at the origin of replication of the H-strand (O_H), in the CR, and it is prematurely terminated in the TAS, promoting the formation of a characteristic triple stranded structure called displacement loop (D-loop), with ~ 700 bp (Brown and

Clayton, 2002; Falkenberg et al., 2007). The occurrence of this structure, associated with an asymmetric replication of mtDNA, leads to an increased exposure to mutations, since it leaves one DNA chain single stranded (Saccone, 2005).

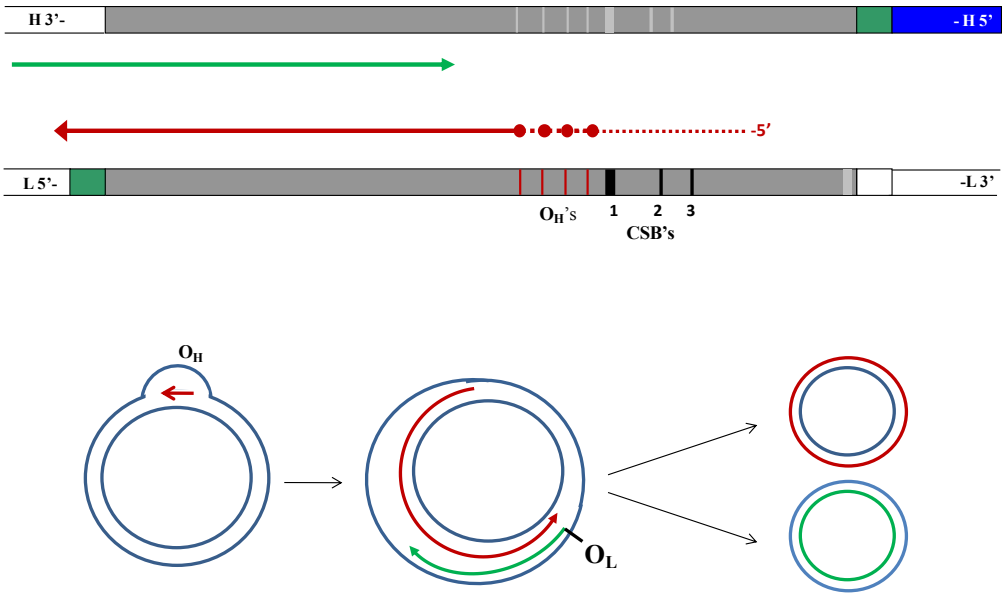


Figure 3 – The strand-asynchronous displacement model for the replication of the mammalian mtDNA. Replication starts unidirectionally on a cluster of sites (O_H) at the control region 3' domain (red line). The RNA primer (red dotted line) required for initiation of H-strand synthesis results from the processing of L-strand transcript. The replication of the L-strand (green line) only starts when the major origin of L-strand DNA replication (O_L) is exposed by the leading strand synthesis.

1.3. MtDNA evolution

The mtDNA sequence evolution is governed by the stringency of structural and functional constraints acting on different genomic regions. For instance, purifying (negative) selection can act against the fixation of nonsynonymous amino acid replacements with a severe effect on the assemblage of complexes of the electron transport chain. A different rate of nucleotide substitutions throughout tRNA and rRNA genes can also result from structural constraints on RNAs secondary structures. The rate of nucleotide substitution is reduced in paired bases due to strong selective pressures to maintain structure-relevant stem regions, while unpaired bases at loop regions usually accumulate more base alterations. The formation of secondary structure in tRNA and rRNA genes at the DNA level during mtDNA replication and transcription can also contribute to this phenomenon. It has

been shown that mutational events are more likely to occur on single-stranded DNA due to, for instance, damage from oxygen radicals, implying that paired bases are better protected from alteration than unpaired ones (Hoede et al., 2006; Wright, 2000).

In addition to base substitutions, a number of polymorphic insertions/deletions (indel) of nucleotides have been observed on rRNA genes when comparing different mammalian species. Interestingly, these polymorphic variants form a peculiar pattern throughout rRNA sequences: indel-rich regions corresponding to loop domains of secondary structures interspersed with highly conserved segments related with paired nucleotides.

A higher-than-expected rate of conservation is also very common in certain mtDNA control region segments due to the presence of preferential sites for the binding of trans-acting nuclear-encoded factors or due to the formation of intra-strand secondary structures with putative regulatory roles (Pereira et al., 2008).

1.4. Heteroplasmy

The presence of a high number of copies of mtDNA in mammalian cells makes it useful in situations in the presence of degraded or low-copy DNA samples. However, the large number of mtDNA molecules have an important consequence in forensic mtDNA profiling: there is the possibility of two or more different mtDNA sequences existing in the same cell or organism, a phenomenon known as heteroplasmy. There are two types of heteroplasmy: sequences may differ at a specific site (point heteroplasmy) or, due to indels (length heteroplasmy) (Andrew et al., 2011; Just et al., 2015). Although the occurrence of heteroplasmy may be an exception to the rule, its high frequency of occurrence in the population means it has to be taken into account in the interpretation of forensic evidence. Thus, even though heteroplasmy can complicate the interpretation of the results, it can actually increase the strength of evidence in some cases, since the occurrence of similar heteroplasmic variation in two different samples by chance could be a rare event (Butler, 2012; Just et al., 2015; Pereira et al., 2010)

Initially, it was thought that heteroplasmy was only present in individuals with diseases, when the frequency of heteroplasmy reached a certain threshold (Ye et al., 2014). Nowadays, with proper procedures for the detection and validation of heteroplasmy, it is possible to observe that heteroplasmy is more common than

previously expected, being also present in healthy individuals (Just et al., 2015; Li et al., 2010; Ramos et al., 2009; Ye et al., 2014).

Although heteroplasmic sites can be found throughout the whole mitogenome, it is more common in the control region (Ramos et al., 2009). Nevertheless, the reason why some mtDNA positions are particularly prone to heteroplasmy remains unknown. A possible explanation for the existence of mutational and heteroplasmic hotspots is the background sequence, which can be prone to the formation of DNA secondary structures, also known as non-B or non-canonical DNA (Pereira et al., 2008).

1.5. Non-B or non-canonical DNA conformation

In 1953, Watson and Crick described the antiparallel, right handed and double helical B-form of DNA, which is now accepted as the canonical form. However, it was soon discovered that DNA could also acquire non-canonical structures, which are abundant in the mammalian nuclear genome (Bacolla et al., 2010, 2013). In fact, non-B conformations occur more frequently than previously expected and associated with specific genome regions. Thus, what was first thought to be junk DNA, it is now known to be related to different processes in the genome (Zhao et al., 2010).

The stability of the DNA molecule is constantly being challenged and it is always in a dynamic state, with transitions between B-form and non-B DNA. Nonetheless, DNA molecule is in its canonical conformation most of the time and the occurrence of secondary structures is only transient and associated with specific sequence motifs (Wang and Vasquez, 2014; Wells, 2007). So far, a large number of non-B DNA conformations have been described, such as hairpins, cruciforms, triplexes, quadruplexes and left-handed Z form, among others (Bacolla et al., 2010, 2013). As it was stated before, certain sequence motifs are necessary in order to occur the formation of such structures (e.g. short tandem repeats, inverted repeats and tracts of $(G)_n$) and it is facilitated by biological processes, such as transcription, replication and DNA repair, that in addition to provide energy from negative supercoiling, also promote the separation of both strands and, thus, while DNA is single stranded (ssDNA) the occurrence of intra-strand folding is possible (Hoede et al., 2006; Wang and Vasquez, 2014).

Non-canonical conformations are known to be associated with genetic instability, which can cause a several genetic human diseases such as myotonic dystrophy, Huntington disease, Friedrich ataxia or fragile X syndrome (Bacolla et al., 2006; Wells, 2007). This instability may be due replication slippage caused by secondary structures during replication, giving rise to expansion or contraction of repeats (indels) and resulting in the rearrangement of the genome (Wang and Vasquez, 2014; Wells, 2007). Another cause of genetic instability is the formation of DNA secondary structures, particularly in single-stranded DNA regions (ssDNA), which is much more prone to mutations than double-stranded DNA (dsDNA) (Wright, 2000). The existence of single- and double-stranded DNA regions occurs with the formation of hairpins or cruciform (**Figure 4**). These structures are formed due to a sequence composed by two inverted repeats segments (i.e. series of bases followed by their complementary on the same strand), separated by a track of noncomplementary nucleotides, allowing it to fold back upon itself, to form the stem, and the sequence in the middle will become the loop at the end of the stem (Bacolla et al., 2010; Wright, 2000). Therefore, unpaired and mispaired bases in secondary structure, such as hairpins, form single-strand loops that are much more vulnerable to mutations than paired bases in double-stranded stem regions (Hoede et al., 2006; Wright, 2000). By this reason, the existence of mutational and heteroplasmic hotspots in mtDNA can be related with the formation of non-B DNA structures.

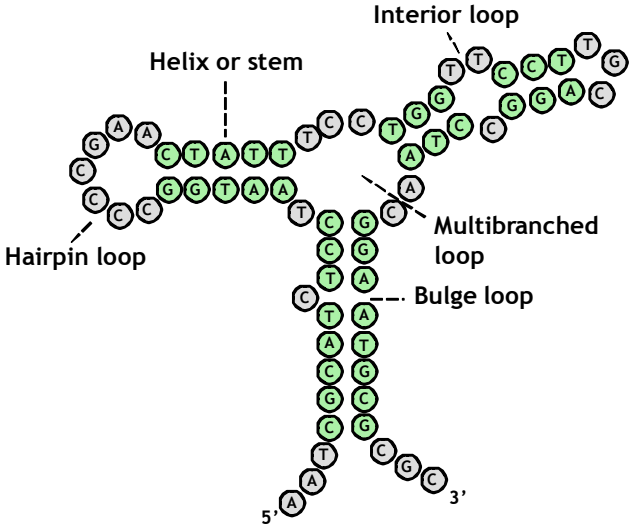


Figure 4 – Schematic representation of a hypothetical DNA secondary structure (paired bases in green and unpaired in gray). A stem region with different types of loops is indicated.

Previous investigations suggested that non-B DNA structures could also occur in the mitochondrial genome, playing an important role in the complex genomic

rearrangements and in mitochondrial diseases (Damas et al., 2012). Actually, the human mtDNA is composed by regions with the potential to form non-canonical conformations (Hou and Wei, 1998), which along with the fact that it is continuously being replicated and transcribed, makes mtDNA prone to form non-canonical structures (Damas et al., 2012).

1.5.1. Structure A

We previously identified 13 non-B DNA structures in the human mtDNA control region (Pereira et al., 2008), with the most striking results observed for a 93-nt cloverleaf-like structure, named structure A (mtDNA reference positions 16028 – 16120), which presented several evidences of *in vivo* formation (**Figure 5**). This structure has 60 paired bases and 33 unpaired and it is under both negative and positive selection, with higher proportion of polymorphic positions observed in unpaired nucleotides when compared with stem regions and the occurrence of compensatory base changes in stem regions. Structure A presented the lowest free energy (ΔG = -11.17 kcal/mol) among all the 13 structures and had one of the highest folding potential of the entire mitogenome.

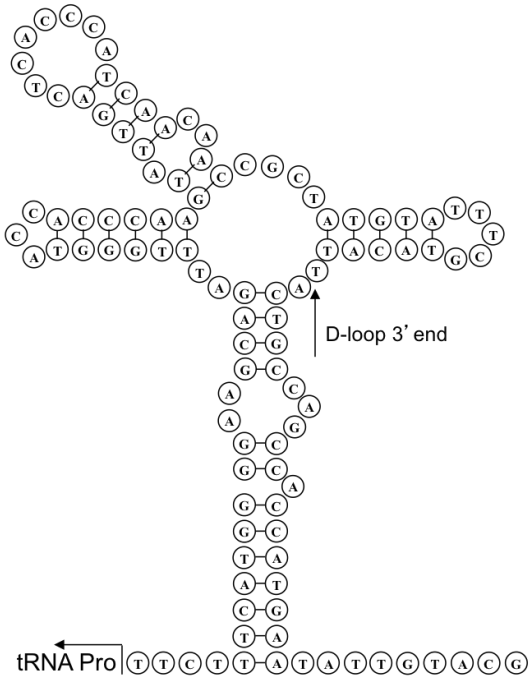


Figure 5 – Schematic representation of the predicted cloverleaf-like structure A, located in the human mtDNA control region. The D-loop 3' end is shown.

The formation of structure A might be responsible for the observed heterogeneity in the rate of substitutions among sites, in addition to be a hot spot for the 3'-end of human mtDNA deletions. Moreover, the fact that structure A is located near the control region 5'-end, comprising the D-loop stop point, suggests it may play an important regulatory role during replication (Pereira et al., 2008). The position 16093 lies within one of the loops formed in the predicted cloverleaf-like structure A and presents a high percentage of heteroplasmy in a wide range of different tissues (Irwin et al., 2009; Just et al., 2015; Krjutskov et al., 2014; Naue et al., 2015; Samuels et al., 2013). It remains to be determined if the formation of structure A is related with the existence of heteroplasmic hotspots, as position 16093.

2. Aims

The main goal of this study is to provide new evidence for the formation of non-B DNA structures in the human mitochondrial genome. In particular, we aimed to:

- Design single and double stranded DNA constructs with the predicted non-B DNA elements from mtDNA structure A;
- Infer the occurrence of hairpins in vitro through endonuclease S1 nuclease digestion in the DNA constructs;
- Visualize non-B DNA elements in the duplex DNA constructs with atomic force microscopy;
- Verify how heteroplasmic positions are distributed in the mtDNA to infer possible modifications in the predicted non-B DNA structures.

3. Material and Methods

3.1. Design of complementary ssDNA oligonucleotides

Two complementary sets of oligonucleotides named “oligo A” and “oligo B” were designed as presented in **Figure 6**.

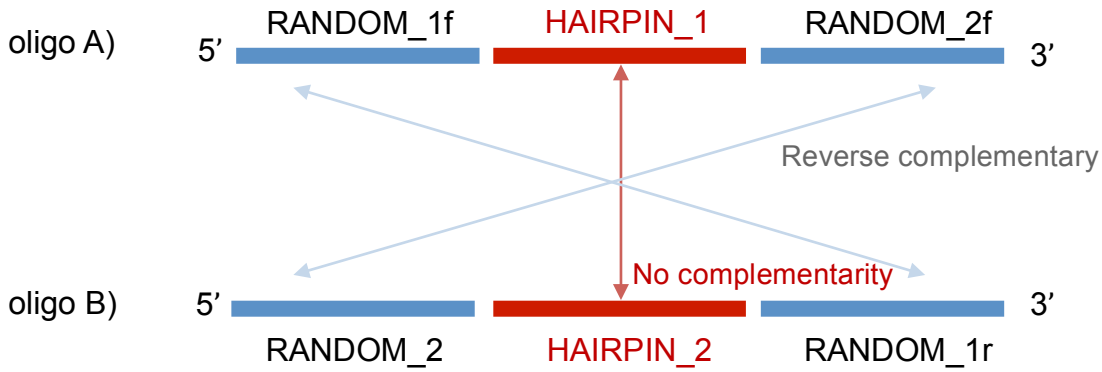


Figure 6 – Schematic representation of two single-stranded oligonucleotides (oligo A and oligo B) used to make DNA duplexes. The random flanking sequences reverse complementary regions are in blue (Random_1f - Random_1r and Random_2f - Random_2r) and the non-complementary sequences with predisposition to form secondary structures in red (Hairpin_1 and Hairpin_2).

Oligo A was designed to include random sequences at the 5’ and 3’ ends (RANDOM_1f and RANDOM_2f) flanking a central sequence (HAIRPIN_1) with a predicted non-B DNA conformation (i.e. hairpins) (Pereira et al., 2008). Regarding HAIRPIN_1, in order to infer if the formation of structure A was possible, the different hairpin sequences that form within the cloverleaf-like structure were used: Hairpin_16045, Hairpin_16060 and Hairpin_16087 (the numbers indicate the reference position where the structure begins) (**Figure 7**).

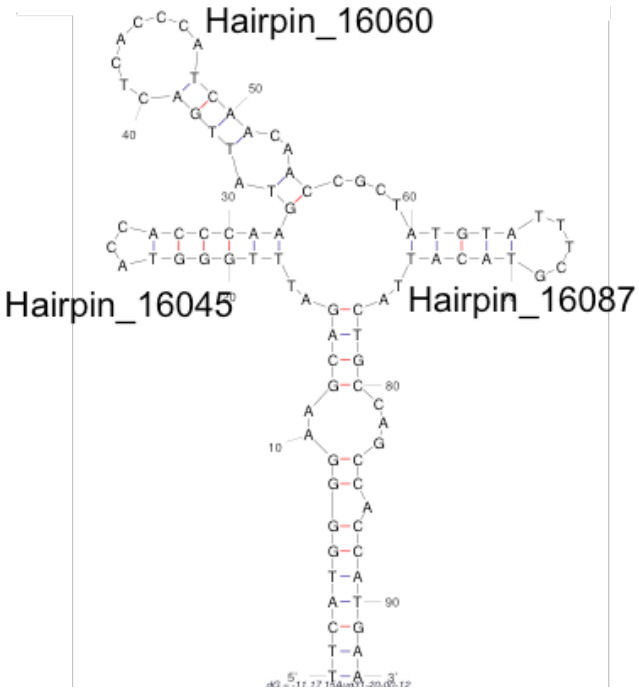


Figure 7 – Graphical representation obtained with mFold of structure A with the names for each hairpin formed in the cloverleaf-like structure (Hairpin_16045, Hairpin_16060 and Hairpin_16087).

Part of the stem sequence of structure A was also used, Hairpin_Stem, intercalated by a poly-T tract in order to form the loop at the end of the stem. As a control for the formation of non-canonical structures in the oligo A, a sequence that includes the origin of replication of the L-strand (O_L) of the human mtDNA was used (Hairpin_O_L) which forms a well known characteristic hairpin (Hixson et al., 1986) (**Figure 8**).

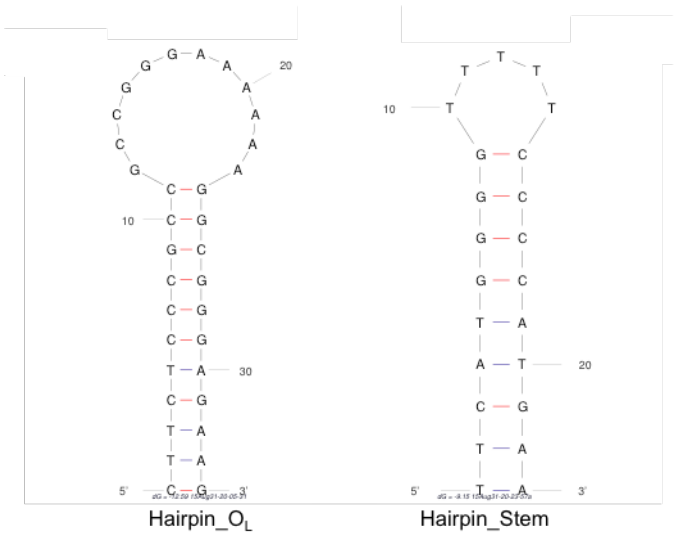


Figure 8 – Graphical representation obtained with mFold of the predicted secondary structures formed by Hairpin_O_L and Hairpin_Stem.

The 20 nucleotides random sequences were selected to avoid self-complementarity or intra strand folding, as determined using the OligoCalc software and confirmed with mFold (version 3.5; http://www.bioinfo.rpi.edu/applications/mfold/), which is a software for DNA secondary structures prediction by free energy minimization using nearest-neighbor thermodynamics rules (Markham and Zuker, 2005; SantaLucia, 1998; Zuker, 2003). The regions were also designed to contain 50% of G/C in order to promote a strong bond between the strands but not to the point of forming secondary structures. Another feature of these random sequences was the inclusion of one A/T base flanking the hairpins, which provides a weaker bond between the two oligos and facilitates the formation of the structure, and two G/C in the end of the oligonucleotide, to achieve a stronger bond between oligo A and oligo B.

Oligo B included two flanking random sequences (RANDOM_2r and RANDOM_1r) reverse complementary to the oligo A random sequences. A 37 nt sequence (AAAGTCCTAGCAATCCAAATGGGATTGCTAGGACCAA) from the ColE1 plasmid was used for the central hairpin (HAIRPIN_2), which is known to adopt a non-B DNA conformation (Lilley and Kemper, 1984). The size of the sequence was then adjusted to 15 nt, 23 nt and 31 nt (Hairpin_ColE1.55, Hairpin_ColE1.63 and Hairpin_ColE1.71, respectively), in order to be similar to the corresponding HAIRPIN_1 sequence size (**Figure 9**). This way, the size of the HAIRPIN_2 sequence will not interfere in the annealing process of the oligonucleotides, allowing the possible formation of secondary structures. The sequences and of the oligonucleotides are presented in **Table 1** and **Table 2**.

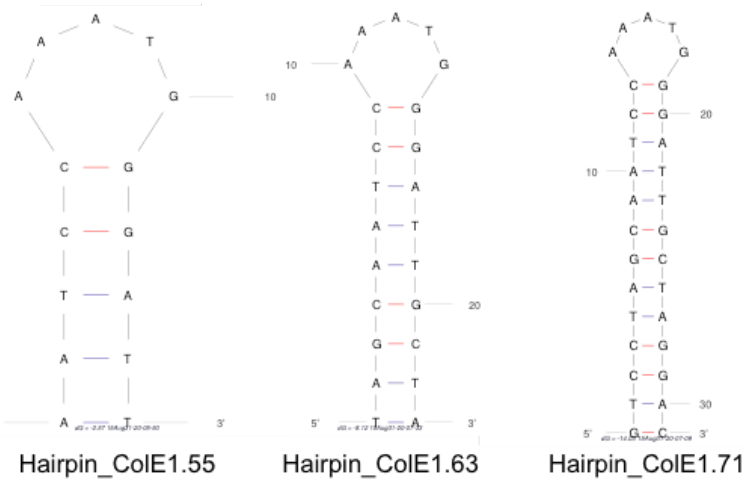


Figure 9 – Graphical representation obtained with mFold of the predicted secondary structures formed by Hairpin_ColE1.55, Hairpin_ColE1.63 and Hairpin_ColE1.71.

Table 1 – Sequence of each region composing Oligo A and Oligo B.

Name	Sequence (5' – 3')
RANDOM_1f	GCTATGACGTTGCTGACTCT
RANDOM_2f	TCGCGAATGGTAAACACTCC
Hairpin_16045	TTGGGTACCACCCAA
Hairpin_16060	GTATTGACTCACCCATCAACAAC
Hairpin_16087	ATGTATTTCGTACAT
Hairpin_O_L	CTTCTCCCGCCGCCGGGAAAAAAGGCGGGAGAAG
Hairpin_Stem	TAGCAATCCAAATGGGATTGCTA
Hairpin_ColE1.55	AATCCAAATGGGATT
Hairpin_ColE1.63	TTCATGGGGTTTTTCCCCATGAA
Hairpin_ColE1.71	GTCCTAGCAATCCAAATGGGATTGCTAGGAC

Table 2 – Final sequence of the oligonucleotides (the region with predisposition to adopt secondary structures is underlined).

Oligonucleotide Name	Sequence (5' – 3')
Oligo_16045	GCTATGACGTTGCTGACTCTTTGGGTAC CACCCAATCGCGAATGGTAAACACTCC
Oligo_16060	GCTATGACGTTGCTGACTCTGTATTGACTCAC CCATCAACAACTCGCGAATGGTAAACACTCC
Oligo_16087	GCTATGACGTTGCTGACTCTATGTATTT CGTACATTCGCGAATGGTAAACACTCC
Oligo_O_L	GCTATGACGTTGCTGACTCTCTTCTCCCGCCGCCGGGA AAAAAGGCGGGAGAAGTCGCGAATGGTAAACACTCC
Oligo_Stem	GCTATGACGTTGCTGACTCTTTCATGGGGTTT TTCCCCATGAATCGCGAATGGTAAACACTCC
Oligo_ColE1.55	GGAGTGTTTACCATTCGCGAAATCCAAA TGGGATTAGAGTCAGCAACGTCATAGC

Table 2 – *continued*

Oligo_ColE1.63	GGAGTGTTTACCATTCGCGATAGCAATCCAAA TGGGATTGCTAAGAGTCAGCAACGTCATAGC
Oligo_ColE1.71	GGAGTGTTTACCATTCGCGAGTCCTAGCAATCCAAA TGGGATTGCTAGGACAGAGTCAGCAACGTCATAGC

3.2. Annealing of oligonucleotides assay

The oligonucleotides (Sigma-Aldrich, United States) were purchased lyophilized with HPLC purification. Each oligonucleotide was resuspended in water to a final concentration of 100 µM and stored at -20 °C. The formation of duplex DNA constructs was made by diluting and mixing oligo A with the corresponding oligo B (**Table 3**), in an eppendorf at a 1:1 molar ratio, to a final concentration of 1 pmol/µL with TEN buffer (10 mM Tris-HCl, 1 mM EDTA and 50 mM NaCl). The eppendorf was capped with parafilm and placed in boiling water, in a large glass beaker on a hotplate, for 2-3 minutes. After that time, the hotplate was turned off and the eppendorf was left in the water overnight, to slowly cool to room temperature. The eppendorf was centrifuged to remove the solution in the cap and the aliquot was stored at -20 °C.

Table 3 – Oligo A oligonucleotide with the corresponding oligo B oligonucleotide and the name of each construct after the annealing was performed.

Oligo A	Oligo B	Construct name
Oligo_16045	Oligo_ColE1.55	Construct_16045
Oligo_16060	Oligo_ColE1.63	Construct_16060
Oligo_16087	Oligo_ColE1.55	Construct_16087
Oligo_O_L	Oligo_ColE1.71	Construct_O_L
Oligo_Stem	Oligo_ColE1.63	Construct_Stem

3.3. Endonuclease digestion assay

The duplex DNA constructs were digested with the endonuclease S1 Nuclease (Thermo Scientific, United States) in order to infer the formation of hairpins. For S1 Nuclease digestion, the following mixture was prepared: ~1 µg of DNA (constructs), 16 µL of 5x reaction buffer for S1 Nuclease, 0.1 µL of S1 Nuclease and dH_2O to a final volume of 80 µL. Then, the mixture was incubated at room temperature for 30 minutes and, after that time, the reaction was stopped by adding 5.33 µL of 0.5 M EDTA and heating at 70 °C for 10 minutes. All the undigested samples went through the same conditions (buffers, temperatures, reagents, etc.), except for the enzyme. The samples were stored at -20 °C. As a control for the digestion of unpaired nucleotides by S1 nuclease, two sequences of mtDNA, one with 123 bp (ND2) and the other with 103 bp (ND3), with low folding potential were amplified by PCR and digested, following the manufacturers instructions. Since these sequences have low folding potential, their amplicons should be perfect duplexes, with no unpaired or mispaired nucleotides. Thus, it should not occur digestion of the amplicons by the endonuclease.

3.4. PCR assay

For the amplification of the constructs by PCR, two different sets of primers were used. The random sequences of the oligonucleotides were used to design the primers – normal primers – (5'-GCTATGACGTTGCTGACTCT and 5'-GGAGTGTTTA CCATTCGCG). The difference between the two set of primers was that in one of them a $(GACT)_5$ tail was added at 5', in order to increase the size of the amplicon by 40 bp (tailed primers). Since we were dealing with small molecules with the potential to form secondary structures, increasing the size of each molecule by 40 bp may facilitate the observation after electrophoresis. To amplify ND2 and ND3 sequences the following primers were used: 5'-TCCCCACCATCATAGCCAC and 5'-GTTCAAAC TGTCATTTTATTTTTACG for ND2 and 5'-CTCCATAAAATTCTTCTTAGT AGC and 5'-TGGCAGGTTAGTTGTTTGTAGG for ND3. The QIAGEN Multiplex PCR kit was used and the following mixture was prepared: 5 µL of the 2x QIAGEN Multiplex PCR Master Mix, 1 µL of each primer at 2 µM, 2 µL of DNA and dH_2O to a final volume of 10 µL. The PCR conditions are presented in the **Table 4**. After the reaction was completed, the samples were stored at -20 °C.

Table 4 – Program conditions of PCR reactions.

Cycle step	Temperature	Time	No. of cycles
Hot Start	95 °C	15’	1
Denaturation	94 °C	30’’	35
Primer annealing	55 °C	1’30’’	
Extension	72 °C	1’	
Final extension	72 °C	10’	1

3.5. Electrophoresis and silver staining assay

After amplification, PCR products were submitted to electrophoresis in a horizontal polyacrylamide gel. The T9C5 gel was obtained by mixing 3 mL of acrylamide solution (25 mL of 4x gel buffer – 1.5 M Tris-HCl buffer (pH 8.8) –, 20 mL of Acrylamide:Bisacrylamide 19:1 (40%), 7 mL of Glycerol and 43 mL of dH_2O), 170 µL of 2.5% Ammonium Persulfate (APS) and 7 µL of TEMED. Glass supports, with one side covered by a hydrophilic gel-bond film, were used in order to obtain a gel with 3 mm of thickness. After loading the samples, two paper strips soaked in buffer (Tris-Glycine) were used at both anode and cathode to allow the horizontal run. In the anode paper strip it was added bromophenol blue dye, allowing monitoring the electrophoresis. The electrophoretic system was submitted to refrigeration (4 °C) and to a voltage of 150 V for 50 minutes. After electrophoresis, the gel was submitted to silver staining. This procedure started with a fixation step of the DNA, with 10% ethanol for 10 minutes, always in agitation, followed by 5 minutes in 1% nitric acid. Then, the gel was washed twice in dH_2O for about 10 seconds each. Afterwards, the gel was placed in 0.2% silver nitrate solution for the coloration step, for 20 minutes, in complete darkness. Finished that time, it was washed again twice in dH_2O. Finally, the revelation step was made with a solution of 0.28 M Sodium Carbonate (Na_2CO_3) and 0.02% formaldehyde. When the DNA bands where visible, the revelation reaction was stopped with 10% acetic acid for 2 minutes. The gel was washed with dH_2O and dried at room temperature.

Capillary electrophoresis was also performed (QIAxcel system). PCR products were run using the DNA High Resolution Gel Cartridge (QIAGEN) under method OM700.mtd for 10 s at 5 kV voltage for sample injection and 700 s and 3 kV

voltage for fragment separation. The results were analyzed using QX Biocalculator Fast Analysis Sofware (QIAGEN).

3.6. AFM assay

After the annealing, the constructs were diluted with ultrapure water (5 PRIME) to a wide range of concentrations (1 ng/µL, 3 ng/µL, 10 ng/µL, 15 ng/µL and 20 ng/µL), in order to optimize the final concentration to use in this assay. Then, the aliquots were stored at 4°C. $MgCl_2$ was added, to a final concentration of 2.5 mM, right before 10 µL of this new mix (construct and $MgCl_2$) was pipetted in a freshly cleaved mica and incubated for 10 minutes. During these 10 minutes, more of this mix was pipetted. Afterwards, the mica surface was washed 2-3 times with ultrapure water and it was dried with airflow and placed in a vacuum chamber for 30 minutes, to remove the excess of water. Finally, the sample was scanned in air using a Veeco Multimode NanoScope Iva operating in tapping mode and the software NanoScope v6.13 was used for image acquisition and analysis.

3.7. Heteroplasmy database

3.7.1. Complete mtDNA

The information of heteroplasmic positions was obtained in published data by Irwin et al., 2009, Li et al., 2010, Ramos et al., 2013 and Samuels et al., 2013, through the analysis of a wide variety of samples, such as blood, hair, bone, brain, lung, heart muscle, heart muscle, bladder, buccal swab, among others. We also used the MITOMAP database (Lott et al., 2013), namely the “Somatic Mutations”, “rRNA/tRNA and Coding & Control Region Mutations” databases, accessed in November 2014. The heteroplasmic sites were numbered according to the rCRS.

In order to determine the distribution of the heteroplasmies among paired and unpaired sites throughout the entire mtDNA, a folding prediction for the single-stranded revised Cambridge reference mtDNA sequence (rCRS, NC_012920) was performed using the hybrid-ss-min core programme of the UNAFold 3.8 software package (Markham and Zuker, 2008). The folding of the single-stranded DNA was carried out at a temperature of 37 °C, sodium concentration of 1 M and magnesium concentration of 0 M. DNA folding was limited to structures where the distance

between the bases starting and terminating the primary stem is equal or below 100 nucleotides. The more stabled predicted structure, indicated by the lowest free energy (ΔG), was used. Moreover, we only considered structures with stems composed by more than 10 nucleotides.

3.7.2. Structure A

Regarding the distribution of heteroplasmic positions in the structure A, the previously obtained information and the predicted structure in Pereira et al., 2008 were used. We used the mFold software in order to infer if each heteroplasmy would cause structural alterations comparing with the previously predicted structure A. For that, we tested the sequence of structure A with each heteroplasmy. For each heteroplasmy, the predicted structure with the lowest ΔG was compared with the predicted structure A.

3.8. Statistical analysis

Fisher's exact test or Chi-square test ($\alpha = 0.05$), dependind on the size of the sample, were performed using GraphPad Prism version 6.01 for Windows, Graphpad Software, San Diego California USA, www.graphpad.com.

4. Results and Discussion

4.1. *In vitro* formation of secondary structure in DNA constructs

Since previous studies were solely based on computer predictions and phylogenetic inferences, our aim here was to demonstrate that mtDNA could form stable non-B DNA structures (i.e. structure A) *in vitro*. Therefore, we designed oligonucleotides with potential to adopt a non-B DNA in order to perform enzymatic reactions with endonuclease S1, which is very specific to detect cruciforms and hairpins formation (Wang et al., 2009). We started by predicting the secondary structure of the designed oligonucleotides Oligo_16045, Oligo_16060, Oligo_16087, Oligo_O_L, Oligo_Stem, Oligo_ColE1.55, Oligo_ColE1.63 and Oligo_ColE1.71 using the mFold software. The predicted sequence with the lowest free energy for each oligonucleotide is presented in **Figure 10**. We used an oligonucleotide when the predicted secondary structure was formed in the region corresponding to HAIRPIN_1 or HAIRPIN_2, with no interaction between these regions and the random flanking sequences.

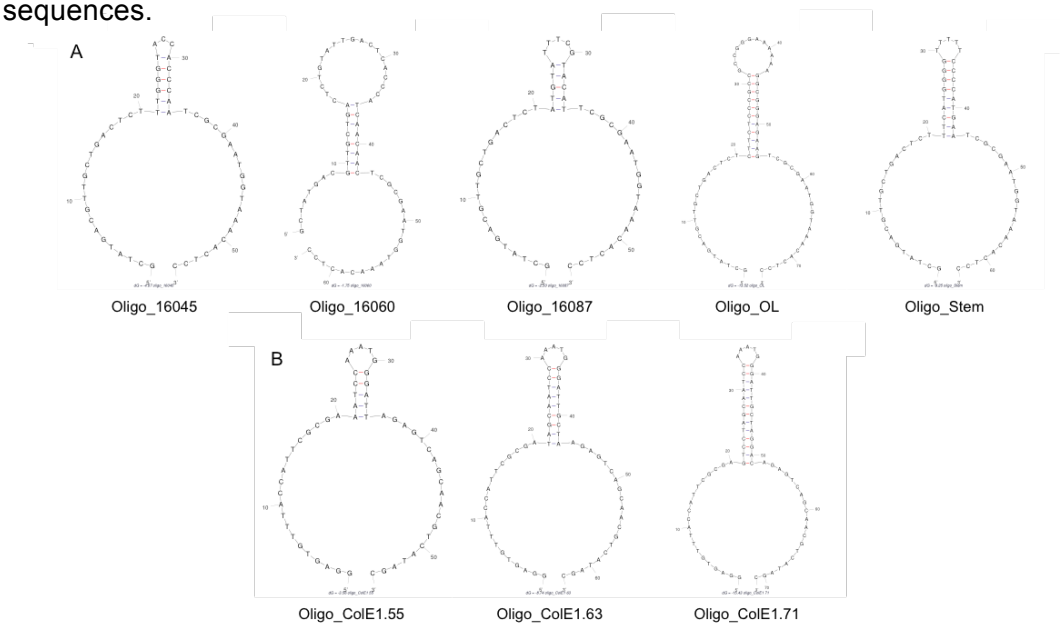


Figure 10 – Graphical representation obtained with mFold of the secondary structure with the lowest free energy in all the designed oligonucleotides. (**A**) Oligonucleotides from oligo A group: Oligo_16045, Oligo_16060, Oligo_16087, Oligo_O_L and Oligo_Stem. (**B**) Oligonucleotides from oligo B group: Oligo_ColE1.55, Oligo_ColE1.63 and Oligo_ColE1.71.

Seven oligonucleotides were predicted to adopt the required secondary structure and were used in further investigations. The Hairpin_16060 was not used since its predicted structure was different from the observed in structure A. The predicted structure of Hairpin_16060 includes the random flanking sequences, instead of only including the central hairpin, which is probably related with the existence of unpaired nucleotides (internal loop) in the stem (**Figure 11**).

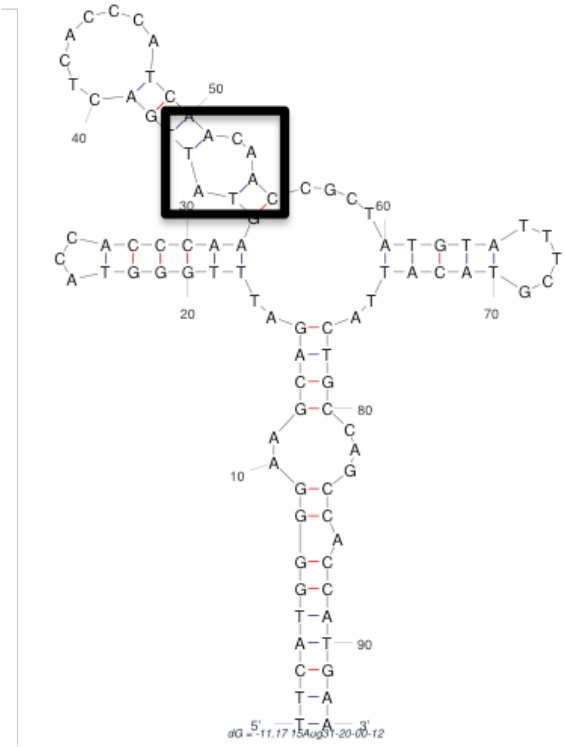


Figure 11 – Graphical representation of structure A obtained with mFold where is highlighted the loop that is formed within the stem of Hairpin_16060.

The predicted structure for Oligo_16087 including the heteroplasmic position 16093 was according to our expectations, allowing it to be used in *in vitro* experiments (**Figure 12**).

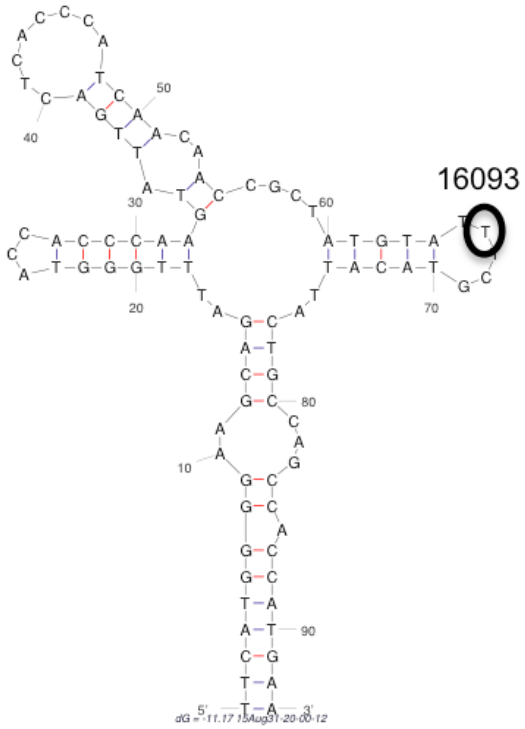


Figure 12 – Graphical representation of structure A obtained with mFold where is highlighted the position of the highly heteroplasmic position 16093, within the loop of Hairpin_16087.

4.2. Endonuclease digestion

4.2.1. Construct_16045

In order to confirm the viability of the procedure to detect non-B DNA structures, a first trial was made using one of the oligonucleotides designed from a hairpin from the structure A (Oligo_16045). We expect the formation of Construct_16045 with a cruciform structure, with the hairpins corresponding to Hairpin_16045 and Hairpin_ColE1.55. Such structure would leave unpaired nucleotides in the loops, which are targeted by nuclease S1 (Lehman, 1981). The cleavage of the construct by endonuclease S1 will prevent the normal progression of the polymerase chain reaction (PCR), meaning that there should be no amplification of the construct and, consequently, no visible bands in the polyacrylamide gel after electrophoresis. In order for the PCR amplification do not occur and, therefore, leading to the absence of bands in the gel, both oligonucleotides that form the construct (in this case Oligo_16045 and Oligo_ColE1.55) had to be cleaved by S1 nuclease. Otherwise, after the second cycle, there would be a dsDNA from the amplified intact strand and consequently normal amplification thereafter (**Figure 13**).

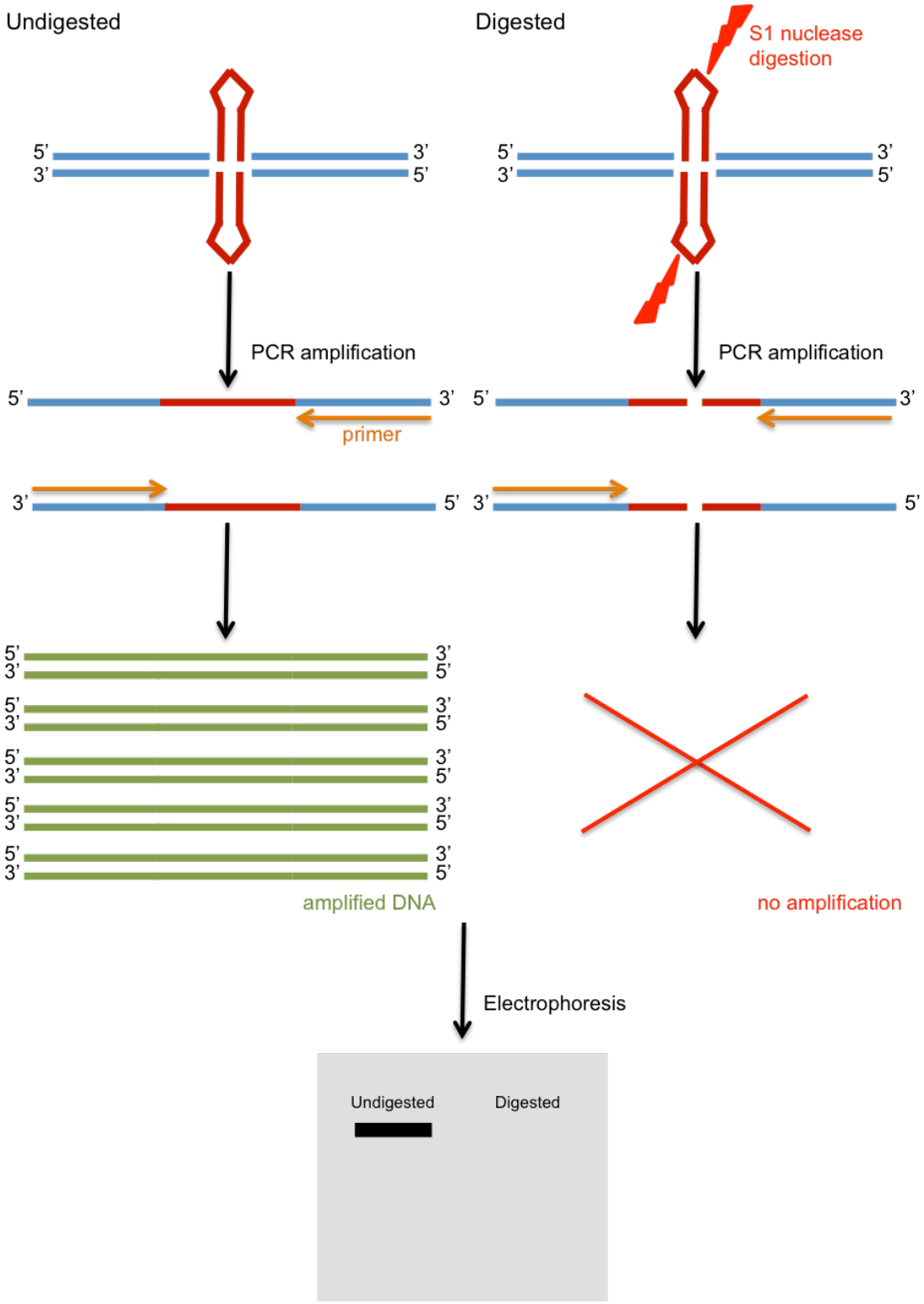


Figure 13 – Schematic representation of the procedure for the *in vitro* detection of ssDNA in the designed constructs. In blue is represented the random sequences from the oligonucleotides, flanking the hairpins sequences (red). The primers are represented in orange, with the arrow pointing to the direction of the synthesis of the new strand. In undigested samples, PCR amplification occurs normally, producing amplified DNA (green) and a visible band in the gel after electrophoresis. In the digested samples, the S1 nuclease will cleave the unpaired nucleotides from the loop (from both strands), disrupting the sequences from the constructs. This way PCR amplification is prevented and there should be no visible bands in the gel.

In this trial, both sets of primers were used and the results are presented in **Figure 14**.

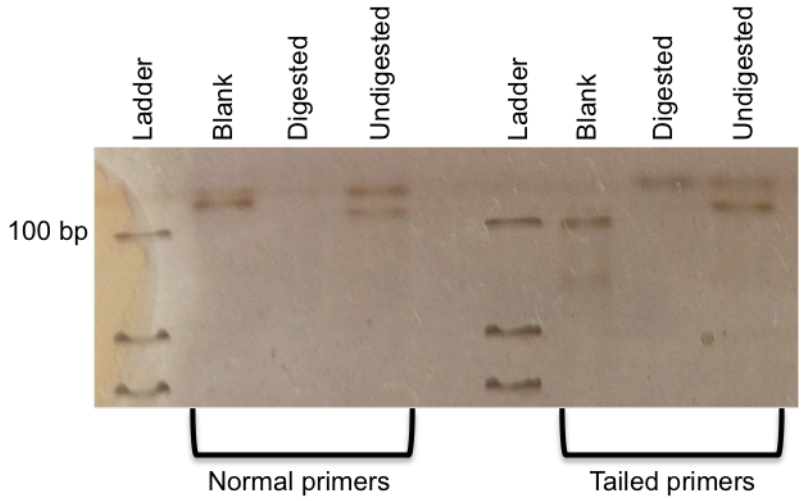


Figure 14 – *In vitro* assay for the detection of ssDNA formation in the Construct_16045, by enzymatic digestion with S1 nuclease. A 100 bp ladder was used to compare with the size of the obtained amplicons with both normal (on the left side) and tailed primers (on the right side).

The digestion with nuclease S1 resulted in no PCR amplification. However, in the undigested sample two amplified products are visible: a shorter amplicon that possibly corresponds to primers dimers, and larger amplicon with the predicted size of the 55 nt construct. This experiment confirmed that the annealing of both strand of the duplex construct occurred as expected, providing evidences for the *in vitro* formation of secondary structures. This result was corroborated by the PCR with tailed primers, from which similar results were obtained. However, in this case we observed a band in the digested sample corresponding to a small amplicons, most likely primer dimers.

4.2.2. Construct_16045, Construct_16087, Construct_Stem and Construct _OL

Then, we tested the remaining constructs (Construct_16045, Construct_16087, Construct_Stem and Construct_O_L) using similar conditions to the trial we made previously with the Construct_16045 (**Figure 15**).

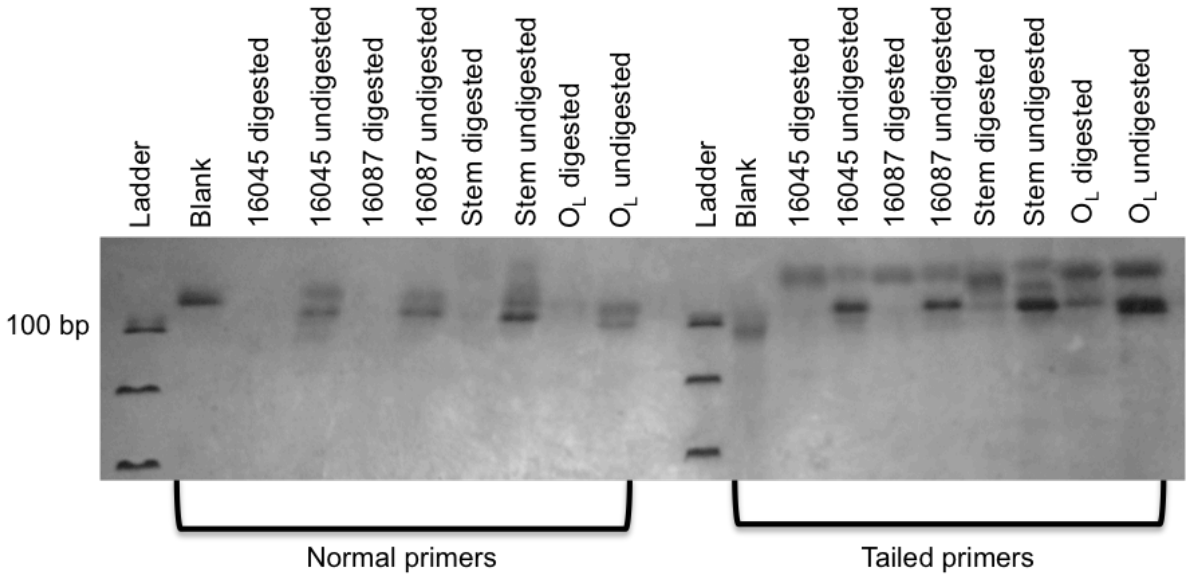
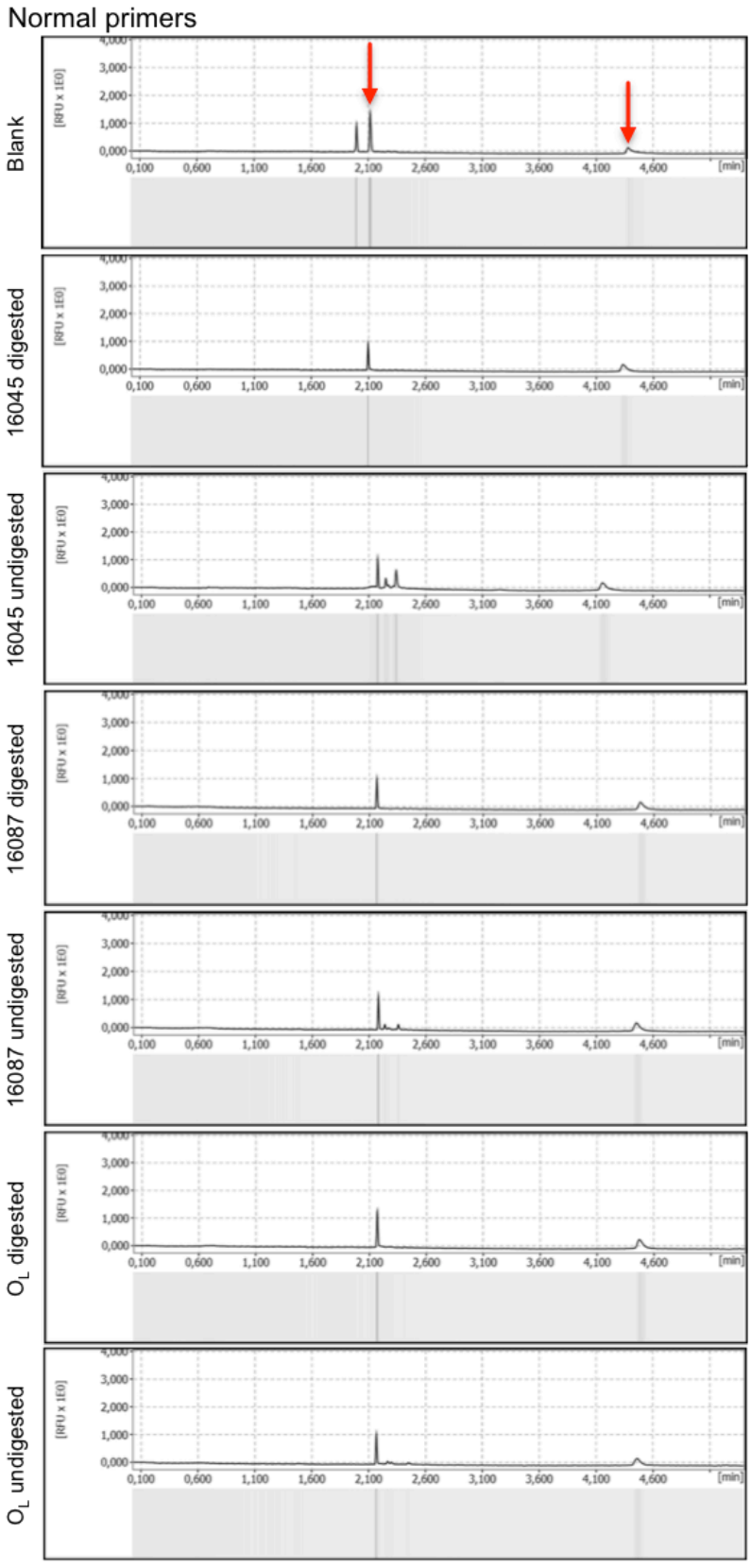


Figure 15 – *In vitro* assay for determination of ssDNA formation in the Construct_16045, Construct_16087, Construct_Stem and Construct_O_L by enzymatic digestion with S1 nuclease. A 100 bp ladder was used to compare with the size of the obtained amplicons with both normal (on the left side) and tailed primers (on the right side).

The results were consistent with those obtained previously with Construct_16045. With both sets of primers, when the constructs were subjected to digestion with S1 endonuclease the normal progression of PCR was compromised and there were no bands present in the gel. The opposite happened in the samples corresponding to the undigested constructs and bands were present in the gel, which means that amplification was able to occur normally. This was also confirmed using QIAxcell (**Figure 16**).

The results obtained with Construct_16045 and Construct_16087 suggest that the sequences of Hairpin_16045 and Hairpin_16087 may have the potential to adopt secondary structures *in vitro*. The results regarding Construct_Stem also point towards the formation of a secondary structure. Altogether, our results suggest that structure A sequences have the potential to form hairpin element *in vitro*. The results obtained with the control sequence known for adopting a non-B DNA conformation (Construct_O_L) were similar to those of structure A elements, which support their formation *in vitro*.


Normal primers
Blank
16045 digested
16045 undigested
16087 digested
16087 undigested
O_L digested
O_L undigested
[RFU x 1E0]
4,000
3,000
2,000
1,000
0,000
0,100
0,600
1,100
1,600
2,100
2,600
3,100
3,600
4,100
4,600
[min]

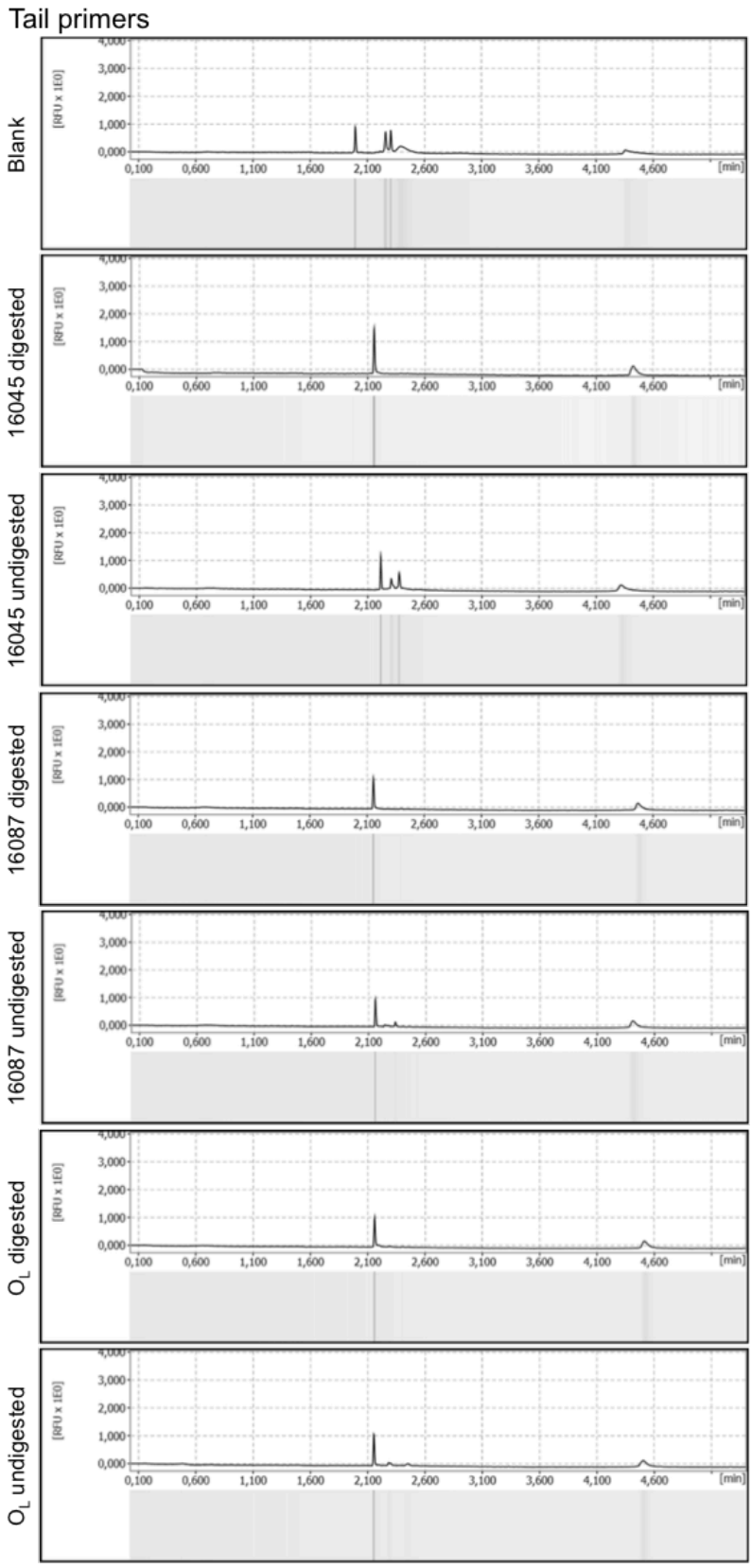


Figure 16 – QIAxcell output for Construct_16045, Construct_16087 and Construct_O_L after amplification using normal and tailed primers. The peaks near 2,100 min and 4,600 min (red arrows) correspond to the reference markers for the determination of the size of the DNA from the samples. The existence of more peaks represents that the sample had DNA. The digested samples had the peaks corresponding to the reference markers, while the undigested also presented peaks corresponding to DNA.

4.2.3. ND2 and ND3

We used two control sequences from mtDNA (ND2 and ND3) with no potential to form secondary structures in order to infer if the absence of bands was specific to constructs with potential to adopt structures when digested with S1 nuclease. These sequences have low folding potential and their amplicons should be perfect duplexes, without unpaired or mispaired nucleotides. Thus, there should be no digestion of the amplicons by the endonuclease and a positive PCR amplification (**Figure 17**).

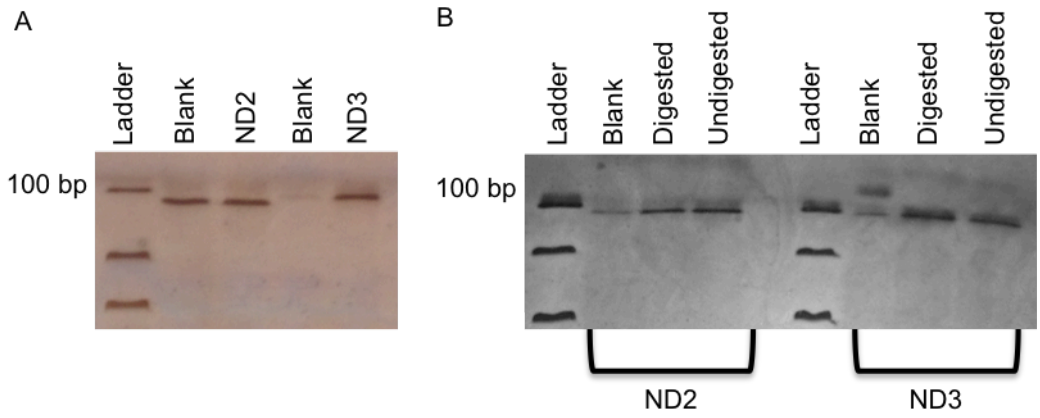


Figure 17 – (**A**) Amplification of ND2 and ND3 sequences from mtDNA for future digestion. (**B**) *In vitro* assay for determination of ssDNA formation in the ND2 (on the lef side) and ND3 (on the right side) sequences, by enzymatic digestion with S1 nuclease. A 100 bp ladder was used to compare with the size of the obtained amplicons.

The two controls were amplified after digestion and bands of expected size were obtained (Figure 17.A). The digested or undigested samples yield bands with the expected size for both ND2 and ND3, suggesting the formation of perfect DNA duplexes (Figure 17.B). These results suggest that the formation of unpaired nucleotides in Construct_16045, Construct_16087, Construct_Stem and Construct_O_L may be related with the formation of secondary structures.

4.3. AFM

We intended to use AFM to try to visualize the secondary structures that would possibly occur in our constructs. AFM consists in a sharp tip scanning through a surface (i.e. mica), pre-treated with $MgCl_2$ to increase the affinity of the surface to DNA, in the x-y axis, reading the profile of the sample and reproducing that surface

profile (z axis). This technique allows directly observation of DNA, appearing as a filament with uniform thickness, and since electron microscopy was not efficient to study the alternative structures, AFM seems attractive to image secondary structures. Moreover, it has already been used in short DNA fragments and to detect secondary structures (Lyubchenko, 2011; Mazur and Maaloum, 2014). We tried to optimize a procedure by testing different concentrations of $MgCl_2$ and DNA, in order to determine the proper concentrations of both to be used in AFM. Nevertheless, after several trials, we were only able to obtain images of a plane surface with what looked like background noise**.** The background noise may be caused by $MgCl_2$ deposits or the DNA constructs, in any case without resolution to visualize any secondary structure (**Figure 18**).

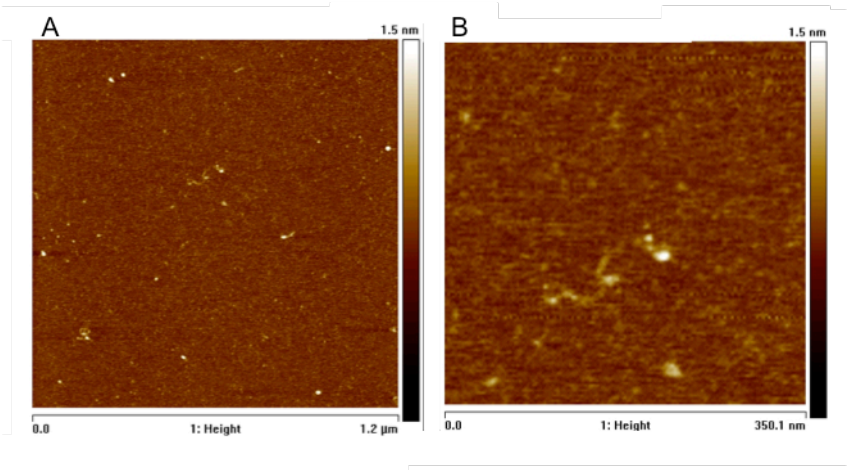


Figure 18 – AFM images of the Construct_16045 acquired in air with a Veeco Multimode NanoScope Iva operated in tapping mode.

The observed “dots” are consistent with DNA deposited in the mica at high concentrations. We concluded that our constructs were too small to a clear observation and discrimination by the available AFM equipment. Although AFM has been successfully used with small DNA molecules (~100 bp), the tested AFM equipment had not enough resolution to observe any structure in our samples. Future experiments should be designed with large constructs, using for example the entire structure A inside a plasmid.

4.4. Distribution of heteroplasmy in the entire mtDNA

We decided to verify how heteroplasmic positions would distribute through the mtDNA. We started by constructing a database with 579 published heteroplasmic sites, each one annotated according to their location in a stem, internal loop, terminal loop or ouside a secondary structure. Then, we decided to analyse the distribution using three approaches:

1. Complete mtDNA vs control region (**Table 5**);
2. Outside of structure vs inside the structure – complete and control region (**Table 6**);
3. Paired nucleotides (stem) vs unpaired nucleotides (internal loop and terminal loop) – inside the structure and in both complete and control region (**Table 7**).

Table 5 – Distribution of normal sites and heteroplasmic sites in the complete mtDNA molecule and in the coding region.

	Normal sites	Heteroplasmic sites
Complete	15990	579
Control Region	947	175

Regarding approach number 1, we noted a total of 579 heteroplasmic positions in the 16569 nucleotides (3.49%) that compose the mtDNA molecule, significantly lower ($p < 0.05$) than the 175 in 1122 nucleotides (15.60%) from the control region. These results are in accordance with the expected and previously reported (Ramos et al., 2009).

Table 6 – Distribution of normal sites and heteroplasmic sites out of structure and in a structure, in both complete mtDNA and control region.

	Complete		Control Region	
	Normal sites	Heteroplasmic sites	Normal sites	Heteroplasmic sites
Out of structure	7103	201	484	72
Structure	8887	378	463	103

As far as approach number 2 is concerned, in both complete mtDNA and control region we obtained statistically significant differences ($p < 0.05$) between the localization of the heteroplasmic sites. Concerning the complete mtDNA, we found 201 heteroplasmic sites in a total of 7304 bases (2.75%), which are not part of any structure, and 378 of 9265 nucleotides (4.08%) that are located in structures. In the case of the control region, we identified 72 heteroplasmic sites located outside of structures, in 556 nucleotides (12.95%), and 103 that are located in structures in a total of 463 bases (18.20%). Our results support the possibility that the occurrence of heteroplasmy can be related with the background sequence, with the formation of secondary structures (Pereira et al., 2008).

Table 7 – Distribution of normal sites and heteroplasmic sites in paired (stem) and unpaired (internal loop and terminal loop) nucleotides, in both complete mtDNA and control region.

	Complete		Control Region	
	Normal sites	Heteroplasmic sites	Normal sites	Heteroplasmic sites
Paired	4707	263	292	72
Unpaired	4180	115	171	31

Finally, in the approach number 3, regarding the control region, although we identified 72 heteroplasmic positions in stem regions, in 364 bases (19.78%), and 31 in 202 nucleotides (15.35%), located in terminal or internal loop, there are no significant differences ($p > 0.05$). However, we obtained statistically significant differences ($p < 0.05$) in the complete mtDNA with 263 heteroplasmic positions located in stems, out of 4970 (5.29%), and 115 in 4295 bases (2.68%) presented in either internal loop or terminal loop.

4.5. Distribution of heteroplasmic positions and structural alterations in the structure A

We wanted to verify how the heteroplasmic positions are distributed through the predicted structure A, in order to see whether there is any bias as far as the localization throughout the structure is concerned, and if each heteroplasmy would cause structural alterations comparing with the predicted structure A. All the obtained

results are presented in **Table 8**. Also, for each one of them, we used the predicted structure obtained with the lowest ΔG.

Table 8 – Heteroplamic sites distribution in the predicted structure A, with the respective structure location, and the effect of the mutation in the structure.

Name	Structure location	ΔG (kcal/mol)	Structural alteration (**Y**es/**N**o)
16028C	Stem	-10.42	Y
16029C	Stem	-10.36	Y
16033A	Stem	-7.79	Y
16034A	Stem	-10.19	Y
16035A	Stem	-10.89	Y
16036A	Stem	-10.89	Y
16043G	Internal loop	-10.69	N
16049A	Stem	-7.84	Y
16053del	Terminal loop	-10.78	Y
16054G	Stem	-10.88	Y
16092C	Terminal loop	-11.91	Y
16093C	Terminal loop	-11.17	N

We found a total of 12 published heteroplasmic positions located in the 93 nucleotides that compose the predicted structure A (**Figure 19**), where 8 of them were located in the stem, in a total of 60 paired bases (13.33%), and 4 in a loop (internal or terminal), in a total of 33 paired bases (12.12%) (**Table 9**) (Pereira et al., 2008).

Table 9 – Distribution of normal sites and heteroplasmic sites in the predicted structure A.

	Normal sites	Heteroplasmic sites
Paired	52	8
Unpaired	29	4

Such as the previously results obtained for the distribution of heteroplasmic sites in paired or unpaired nucleotides in the control region, there is no statistically significant difference between the heteroplasmic sites located in the stem and in the internal or terminal loops, in the structure A.

Only 16092C and 16093C obtained structures with a ΔG lower than the obtained with in the structure A and, therefore, the possibility of formation of more

stable structures. However, from these two, only 16093C did not cause changes in the structure, when comparing with the structure A, along with 16043G.

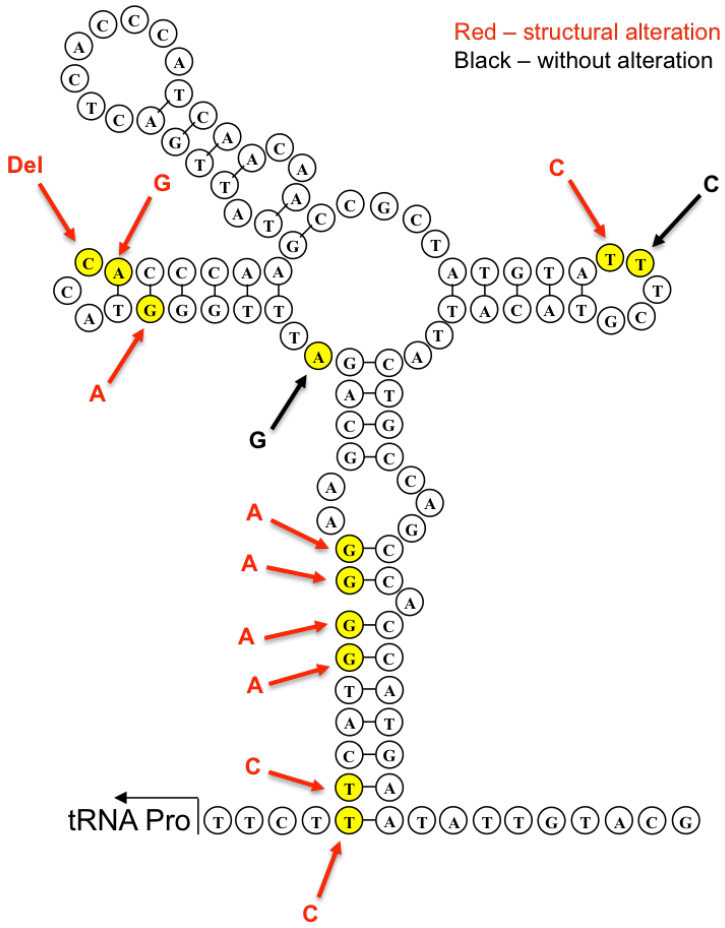
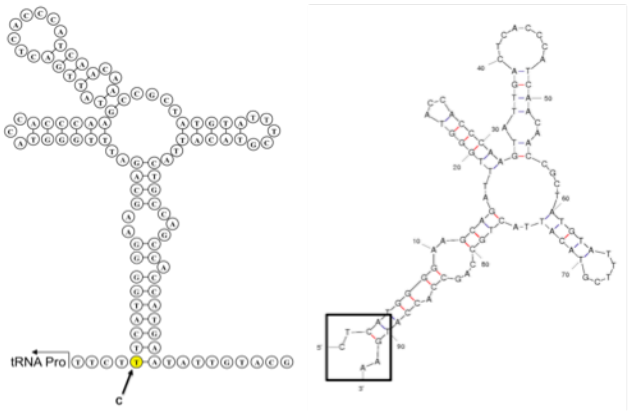


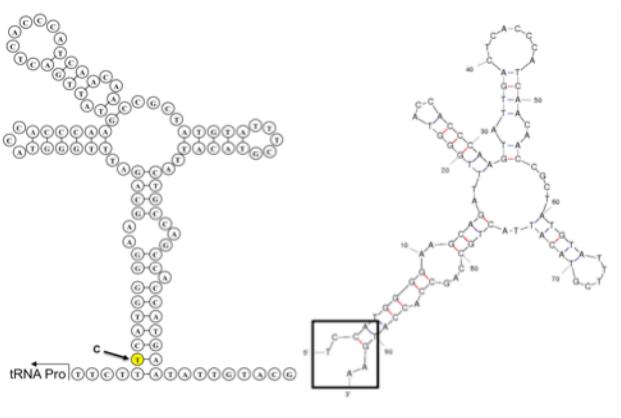
Figure 19 – Schematical representation of the predicted structure A, with the heteroplasmic sites noted (yellow) with the corresponding mutation. In red mutations that cause structural alterations, compared to structure A, and in black mutations that cause no alteration to the structure.

The structural alterations caused by each heteroplasmy, when comparing with the predicted structure A, are presented in **Figure 20**.

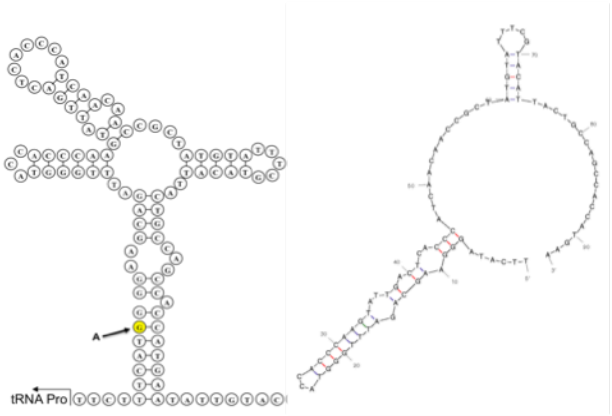
16028C



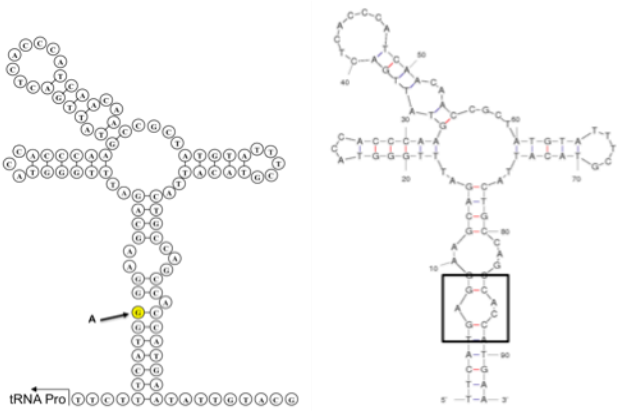
16029C



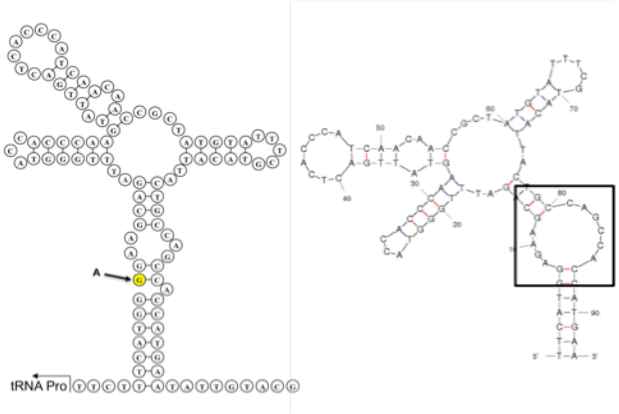
16033A



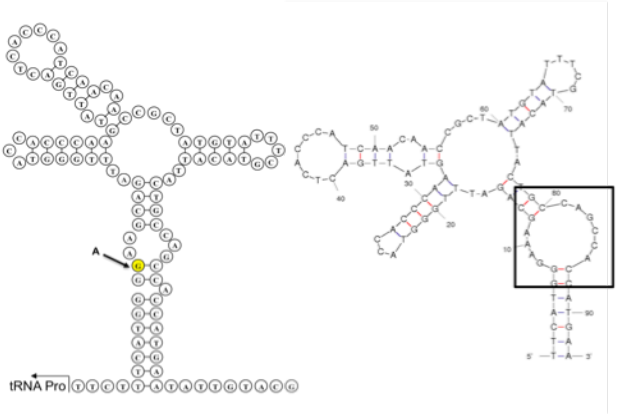
16034A



16035A



16036A



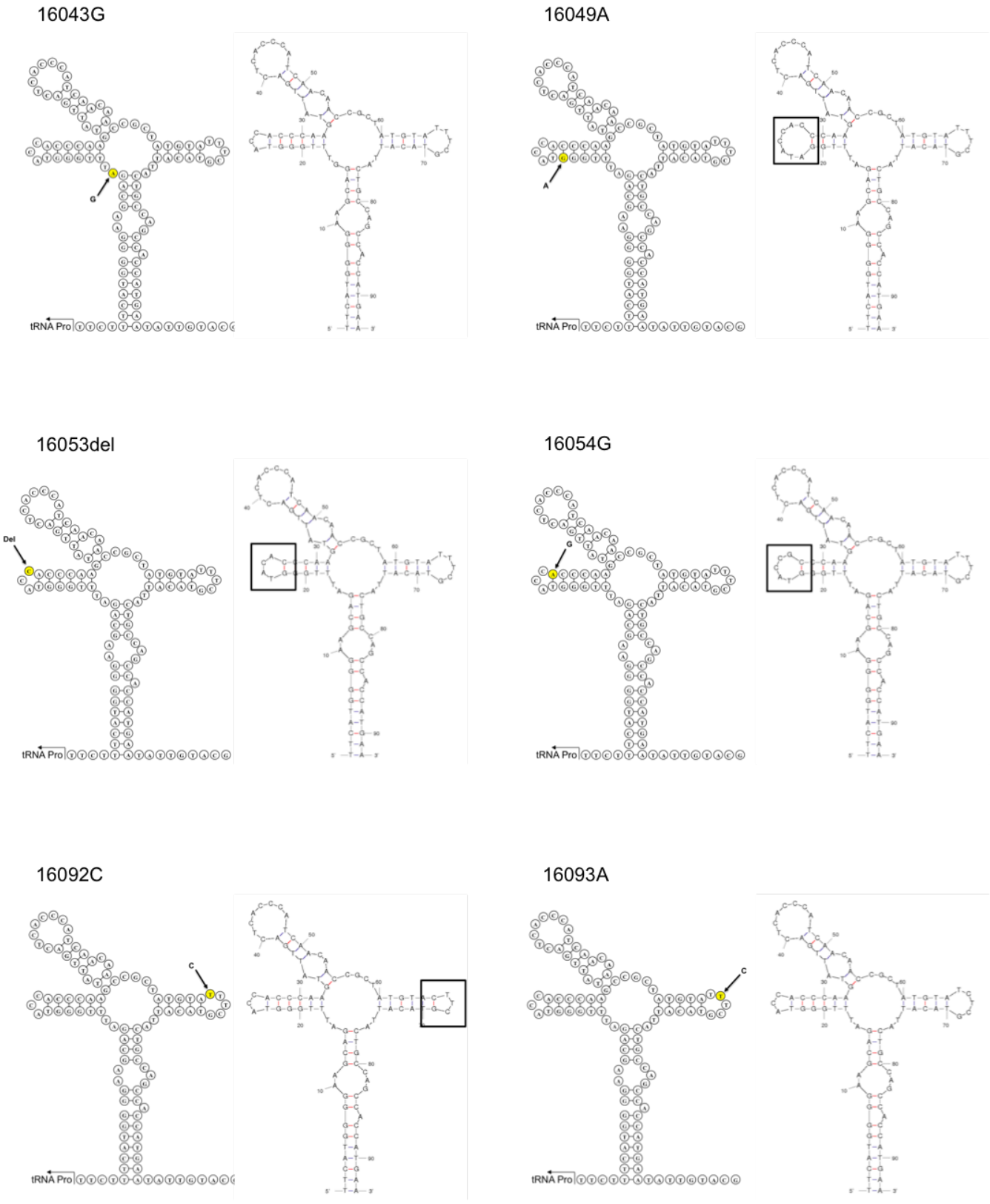


Figure 20 – Graphical representation of the predicted structure A, with the heteroplasmic site noted (yellow) with the corresponding mutation, and the structure with the lowest ΔG obtained for each heteroplasmic position. The square marks the location of the structural alteration. In the case of 16033A there is no square, because the alteration affects all the structure.

As it can be confirmed in Figure 20, only 16043G and 16093C did not cause any structural alterations, when compared to predicted structure A, and they are located in an internal loop and in a terminal loop, respectively. In general, the structural alterations caused by the heteroplasmy were minimal, creating a loop or either stretching or tightening the size of the loop. Nevertheless, the pathological effects that they can lead to still remain unknown. The heteroplasmy 16033A was the only one that obtained, with the lowest free energy (ΔG = -7.79 kcal/mol), a structure with gross alterations comparing with structure A. However, the structure obtained with the second lowest free energy (ΔG = -7.21 kcal/mol), for this heteroplasmy, was more similar to the structure A (**Figure 21**). Nevertheless, the majority of stem of this structure was compromised.

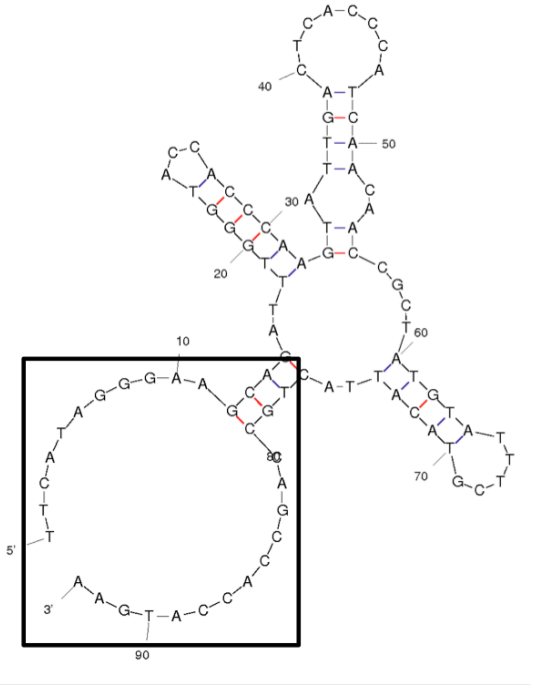


Figure 21 – Graphical representation obtained with mFold of the structure with the second lowest free energy obtained with 16033A. The square indicates the location with the structural alteration, comparing to the structure A.

Regarding 16093C, it is important to note the fact that this mutation does not cause any alteration to the structure. Furthermore, position 16093 has been repeatdly reported, in a wide range of tissues, as a site with high percentage of heteroplasmy, with the mutated base reaching, for example, percentages more than 50% in some of the tissues (Irwin et al., 2009; Krjutskov et al., 2014; Naue et al., 2015). The reason for this heteroplasmic position to be so frequent could possible be the fact that the structure A is not modified with the mutation that occurs, with the transition from thymine to cytocine, and, thus, keeping intact any potential regulatory

role that this structure may play (Pereira et al., 2008). In contrast, the remaining mutations that cause structural alterations may be observed in lower percentages, because possibly, if they reach a certain threshold, they could have an unknown pathological effect.

5. Conclusion

In this study we demonstrate that two of the three hairpins that constitute the cloverleaf-like structure (Hairpin_16045 and Hairpin_16087) have the potential to adopt secondary structure *in vitro*. Unfortunately, we could not test the remaining hairpin from the structure A (Hairpin_16060), because its sequence interacted with the rest of the oligonucleotide sequence and formed a predicted secondary structure different from the expected. Furthermore, we also demonstrated using the Construct_Stem that the stem of structure A has the potential to be formed *in vitro*. This result was supported by the experiment using the controls, Construct_O_L, ND2 and ND3. Altogether, the results obtained supported the possible formation of structure A *in vitro*. In the future, it would be interesting to perform and enzymatic digestion using a restriction enzyme that targets the stem of structure A, which will allow to infer if this structure is really formed, and perform the endonuclease digestion in the entire structure A, although it might be really complicated due to its complex formation.

As far as AFM was concerned, we were unable to optimize a procedure for observation of our constructs, since we did not have enough resolution in the system that we used to observe our constructs. Nevertheless, our results could be useful to devise a different approach in the future. For example, we should increase the size of our target using a plasmid comprising the entire structure and a sharper tip probe and this way maybe we should have no problems with the resolution.

Concerning the distribution of the 579 heteroplasmic positions in the mitogenome, we verified that the control region had more heteroplasmic positions than the complete mtDNA. Moreover, heteroplasmy is more common within predicted structures than outside and the paired nucleotides that compose the stems are more prone to heteroplasmy than unpaired nucleotides.

Finally, we found a total of 12 heteroplasmic positions that are located in the predicted structure A. However, we did not find any statistical significant difference between the heteroplasmic sites located in the stem and in the internal or terminal loop. Ten out of 12 heteroplamies caused structural alterations in structure A. Only 16043G and 16093C did not cause any alteration, keeping the integrity of the structure. Position 16093 has been repeatdly reported with a high percentage of heteroplasmy in a wide range of the tissues, which may be due to the fact that the

structure A is not modified with the mutation that occurs, keeping intact any potential regulatory role that this structure may play. In contrast, the remaining mutations that cause structural alterations may be observed in lower percentages, because possibly, if they reach a certain threshold, they could have any unknown pathological effect.

6. References

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F.*, et al.* (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457-465.

Andrew, T., Calloway, C.D., Stuart, S., Lee, S.H., Gill, R., Clement, G., Chowienczyk, P., Spector, T.D., and Valdes, A.M. (2011). A twin study of mitochondrial DNA polymorphisms shows that heteroplasmy at multiple sites is associated with mtDNA variant 16093 but not with zygosity. PloS one *6*, e22332.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature genetics *23*, 147.

Bacolla, A., Cooper, D.N., and Vasquez, K.M. (2010). Non-B DNA structure and mutations causing human genetic disease. eLS.

Bacolla, A., Cooper, D.N., and Vasquez, K.M. (2013). DNA structure matters. Genome Medicine *5*, 51.

Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J.E., and Wells, R.D. (2006). The involvement of non-B DNA structures in gross chromosomal rearrangements. DNA repair *5*, 1161-1170.

Bernt, M., Braband, A., Schierwater, B., and Stadler, P.F. (2013a). Genetic aspects of mitochondrial genome evolution. Molecular phylogenetics and evolution *69*, 328-338.

Bernt, M., Machné, R., Sahyoun, A.H., Middendorf, M., and Stadler, P.F. (2013b). Mitochondrial Genome Evolution. eLS.

Brown, T.A., and Clayton, D.A. (2002). Release of replication termination controls mitochondrial DNA copy number after depletion with 2',3'-dideoxycytidine. Nucleic acids research *30*, 2004-2010.

Brown, W.M., George, M., Jr., and Wilson, A.C. (1979). Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America *76*, 1967-1971.

Butler, J.M. (2012). Mitochondrial DNA Analysis. In Advanced Topics in Forensic DNA Typing: Methodology (Elsevier Inc), pp. 405-456.

Chen, X.J., and Butow, R.A. (2005). The organization and inheritance of the mitochondrial genome. Nature reviews Genetics *6*, 815-825.

Damas, J., Carneiro, J., Goncalves, J., Stewart, J.B., Samuels, D.C., Amorim, A., and Pereira, F. (2012). Mitochondrial DNA deletions are associated with non-B DNA conformations. Nucleic acids research *40*, 7606-7621.

Falkenberg, M., Larsson, N.G., and Gustafsson, C.M. (2007). DNA replication and transcription in mammalian mitochondria. Annual review of biochemistry *76*, 679-699.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial evolution. Science *283*, 1476-1481.

Hixson, J.E., Wong, T.W., and Clayton, D.A. (1986). Both the conserved stem-loop and divergent 5'-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. The Journal of biological chemistry *261*, 2384-2390.

Hoede, C., Denamur, E., and Tenaillon, O. (2006). Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS genetics *2*, e176.

Hou, J.H., and Wei, Y.H. (1998). AT-rich sequences flanking the 5'-end breakpoint of the 4977-bp deletion of human mitochondrial DNA are located between two bent-inducing DNA sequences that assume distorted structure in organello. Mutation research *403*, 75-84.

Irwin, J.A., Saunier, J.L., Niederstatter, H., Strouss, K.M., Sturk, K.A., Diegoli, T.M., Brandstatter, A., Parson, W., and Parsons, T.J. (2009). Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. Journal of molecular evolution *68*, 516-527.

Just, R.S., Irwin, J.A., and Parson, W. (2015). Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic science international Genetics *18*, 131-139.

Krjutskov, K., Koltsina, M., Grand, K., Vosa, U., Sauk, M., Tonisson, N., and Salumets, A. (2014). Tissue-specific mitochondrial heteroplasmy at position 16,093 within the same individual. Current genetics *60*, 11-16.

Legros, F., Malka, F., Frachon, P., Lombes, A., and Rojo, M. (2004). Organization and dynamics of human mitochondrial DNA. Journal of cell science *117*, 2653-2662.

Lehman, R.I. (1981). Endonucleases specific for single-stranded. In The Enzymes, P.D. Boyer, ed., pp. 193-201.

Li, M., Schonberg, A., Schaefer, M., Schroeder, R., Nasidze, I., and Stoneking, M. (2010). Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. American journal of human genetics *87*, 237-249.

Lilley, D.M., and Kemper, B. (1984). Cruciform-resolvase interactions in supercoiled DNA. Cell *36*, 413-422.

Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D.C. (2013). mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *1*, 1 23 21-21 23 26.

Lyubchenko, Y.L. (2011). Preparation of DNA and nucleoprotein samples for AFM imaging. Micron *42*, 196-206.

Markham, N.R., and Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. Nucleic acids research *33*, W577-581.

Markham, N.R., and Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. Methods in molecular biology *453*, 3-31.

Mazur, A.K., and Maaloum, M. (2014). Atomic force microscopy study of DNA flexibility on short length scales: smooth bending versus kinking. Nucleic acids research *42*, 14006-14012.

McBride, H.M., Neuspiel, M., and Wasiak, S. (2006). Mitochondria: more than just a powerhouse. Current biology : CB *16*, R551-560.

Naue, J., Horer, S., Sanger, T., Strobl, C., Hatzer-Grubwieser, P., Parson, W., and Lutz-Bonengel, S. (2015). Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA. Mitochondrion *20*, 82-94.

Pereira, F., Carneiro, J., and van Asch, B. (2010). A guide for mitochondrial DNA analysis in non-human forensic investigations. The Open Forensic Science Journal *3*, 33-44.

Pereira, F., Soares, P., Carneiro, J., Pereira, L., Richards, M.B., Samuels, D.C., and Amorim, A. (2008). Evidence for variable selective pressures at a large secondary structure of the human mitochondrial DNA control region. Molecular biology and evolution *25*, 2759-2770.

Ramos, A., Santos, C., Alvarez, L., Nogues, R., and Aluja, M.P. (2009). Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification. Electrophoresis *30*, 1587-1593.

Saccone, C. (2005). Mitochondrial Genome. eLS.

Samuels, D.C., Li, C., Li, B., Song, Z., Torstenson, E., Boyd Clay, H., Rokas, A., Thornton-Wells, T.A., Moore, J.H., Hughes, T.M.*, et al.* (2013). Recurrent tissue-specific mtDNA mutations are common in humans. PLoS genetics *9*, e1003929.

SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences of the United States of America *95*, 1460-1465.

Sbisa, E., Tanzariello, F., Reyes, A., Pesole, G., and Saccone, C. (1997). Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. Gene *205*, 125-140.

Taanman, J.-W., and Williams, S. (2005). The human mitochondrial genome: mechanisms of expression and maintenance. In Mitochondria in Health and Disease, C.D. Berdanier, ed. (CRC Press/Taylor & Francis Group: Boca Raton.), pp. 95 - 246.

Thorsness, P.E., and Hanekamp, T. (2001). Mitochondria: Origin. eLS.

Walberg, M.W., and Clayton, D.A. (1981). Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic acids research *9*, 5411-5421.

Wang, G., and Vasquez, K.M. (2014). Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. DNA repair *19*, 143-151.

Wang, G., Zhao, J., and Vasquez, K.M. (2009). Methods to determine DNA structural alterations and genetic instability. Methods *48*, 54-62.

Wells, R.D. (2007). Non-B DNA conformations, mutagenesis and disease. Trends in biochemical sciences *32*, 271-278.

Wright, B.E. (2000). A biochemical mechanism for nonrandom mutations and evolution. Journal of bacteriology *182*, 2993-3001.

Ye, K., Lu, J., Ma, F., Keinan, A., and Gu, Z. (2014). Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. Proceedings of the National Academy of Sciences of the United States of America *111*, 10654-10659.

Zhao, J., Bacolla, A., Wang, G., and Vasquez, K.M. (2010). Non-B DNA structure-induced genetic instability and evolution. Cellular and molecular life sciences : CMLS *67*, 43-62.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research *31*, 3406-3415.